DELTA PUBLICACIONES

100 PROBLEMAS RESUELTOS DE ESTADÍSTICA MULTIVARIANTE

(IMPLEMENTADOS EN MATLAB)

Amparo Baíllo Moreno • Aurea Grané Chávez

100 PROBLEMAS RESUELTOS DE
ESTADÍSTICA MULTIVARIANTE
(IMPLEMENTADOS EN MATLAB)

Acerca de las autoras

Amparo Baíllo Moreno es licenciada y doctora en Matemáticas por la Universidad Autónoma de Madrid, donde trabaja actualmente como investigadora postdoctoral del programa SIMUMAT financiado por la Comunidad de Madrid. Posee un máster en Finanzas Cuantitativas por la Escuela de Finanzas Aplicadas y ha trabajado en el área de Riesgos del Grupo Santander. Cuenta con varias publicaciones científicas en revistas internacionales de impacto y ha participado en distintos proyectos de I+D financiados en convocatorias públicas nacionales. Desde 1998 ha impartido docencia en las universidades Autónoma de Madrid y Carlos III de Madrid.

Aurea Grané Chávez es licenciada y doctora en Matemáticas por la Universidad de Barcelona. Forma parte del Grupo de Análisis Multivariante y Clasificación, vinculado a la SEIO. Cuenta con varias publicaciones científicas en revistas internacionales de impacto y ha participado en distintos proyectos de I+D financiados por la Generalitat de Catalunya y en convocatorias públicas nacionales. En 1994 empezó a impartir docencia en el Departamento de Estadística de la Universidad de Barcelona y actualmente es profesora del Departamento de Estadística de la Universidad Carlos III de Madrid, donde imparte la asignatura Estadística Multivariante en la Diplomatura de Estadística.

100 PROBLEMAS RESUELTOS DE ESTADÍSTICA MULTIVARIANTE

(IMPLEMENTADOS EN MATLAB)

AMPARO BAILLO MORENO
Facultad de Ciencias
UNIVERSIDAD AUTÓNOMA DE MADRID

AUREA GRANÉ CHÁVEZ
Facultad de Ciencias Jurídicas y Sociales
UNIVERSIDAD CARLOS III DE MADRID

DELTA PUBLICACIONES

100 EJERCICIOS RESUELTOS DE
ESTADÍSTICA MULTIVARIANTE
(IMPLEMENTADOS EN MATLAB)

AMPARO BAILLO MORENO
AUREA GRANÉ CHÁVEZ

Editor gerente Fernando M. García Tomé
Diseño de cubierta Mizar Publicidad, S.L.
Preimpresión Delta Publicaciones
Impresión Jacaryan
Avda. Pedro Díez, 3. Madrid (España)

ISBN 84-96477-73-8
Depósito Legal

(0907-60)

A Manolo y Pep

Presentación

El análisis estadístico multivariante es una herramienta de investigación y generación de conocimiento extraordinariamente valiosa, tanto en las ciencias naturales como en las ciencias sociales. Este libro es una valiosa aportación a la literatura en español sobre este tema. Muchos de los interesantes problemas que contiene ayudan a comprender y apreciar el potencial de las técnicas clásicas de análisis multivariante, mientras que otros guían al lector para profundizar en aspectos metodológicos de interés de las técnicas estudiadas. Un atractivo especial de este libro es la inclusión de numerosas rutinas de Matlab que permiten aplicar de forma fácil y flexible las técnicas consideradas a distintos conjuntos de datos reales. Las autoras, Amparo Baíllo y Aurea Grané, tienen gran experiencia en la enseñanza de estas técnicas y el libro muestra claramente su gran experiencia en el análisis de datos reales y en la presentación de los resultados del análisis.

Recomiendo este libro a todos los interesados en las aplicaciones del análisis multivariante y, muy especialmente, a las personas que deseen disponer de un lenguaje potente y flexible, como Matlab, que les permita escribir sus propias rutinas de programación, liberándose del esquema rígido de los programas convencionales. Estoy seguro de que encontrarán este libro muy útil para este objetivo.

Daniel Peña
Catedrático de Estadística
Universidad Carlos III de Madrid

Introducción

El objetivo de este libro es ayudar a comprender todo un conjunto de técnicas exploratorias y estadísticas que permiten sintetizar, representar e interpretar los datos obtenidos de la observación simultánea de varias variables estadísticas. Así pues el libro se centra en el análisis estadístico de matrices de datos, con el fin de extraer de forma rápida la información más relevante contenida en ellas. Los datos de tipo multivariado aparecen actualmente en contextos muy diversos, como son el mundo de la Economía y las Finanzas, las Ciencias Experimentales y la Ingeniería o también en las Ciencias Humanas y Sociales.
Los temas que se tratan pueden clasificarse en tres apartados:

- Inferencia multivariante.
- Técnicas de representación y de reducción de la dimensión.
- Técnicas de clasificación: análisis de conglomerados y análisis discriminante.

Los problemas intentan recoger la diversidad de los campos de aplicación mencionados anteriormente y, en este sentido, se ha procurado buscar conjuntos de datos que fueran interesantes para un público de procedencia muy diversa.
Este libro es fruto de las experiencias docentes de las autoras en la Diplomatura en Estadística y la Licenciatura en Administración y Dirección de Empresas de la Universidad Carlos III de Madrid y en la Diplomatura en Estadística, la Licenciatura en Matemáticas y la Licenciatura en Biología de la Universidad de Barcelona. En general, este libro está dirigido a estudiantes y docentes de cualquier disciplina en la que sea necesario extraer información de un conjunto de datos multivariantes.
Para un seguimiento adecuado del libro se requieren conocimientos básicos de Cálculo de Probabilidades y de Inferencia Estadística. Además son deseables buenos conocimientos de álgebra lineal, más allá de la resolución de sistemas de ecuaciones lineales o de un leve contacto con formas cuadráticas en el contexto del cálculo de extremos de una función real de varias variables. Es quizá demasiado suponer este conocimiento previo y por ello se añade un tema adicional necesario para el desarrollo del libro.

Este libro consta de nueve capítulos. Los tres primeros son introductorios y están dedicados, respectivamente, a una ampliación de conceptos de álgebra lineal, a familiarizarse con las matrices de datos y una introducción a la inferencia normal multivariante. El resto de capítulos están dedicados al estudio de técnicas multivariantes clásicas, como son: el análisis de componentes principales, el escalado multidimensional, el análisis de conglomerados, el análisis factorial, el análisis canónico de poblaciones y el análisis discriminante.

Soporte informático

El volumen de cálculo requerido para el análisis de datos multivariantes hace impracticable su realización manual, no sólo para los cálculos con datos reales, sino incluso si se trata de ejemplos sencillos con datos simulados que ilustren y motiven los conceptos teóricos.

Ya desde los años 70, coincidiendo con la evolución de los ordenadores y la aparición de los primeros paquetes comerciales de programas de Estadística (SPSS, BMDP, SAS), algunos de los autores de libros dedicados al Análisis Multivariante, conscientes de esta situación, han incluido listados de programas para realizar los cálculos correspondientes a las técnicas expuestas.

Por ello hemos creído conveniente disponer de un *software* que permita programar de forma muy sencilla las técnicas que el usuario desea implementar. Esto es posible a través de programas comerciales como MATLAB[1] y S-Plus, o bien sus clónicos gratuitos como OCTAVE y R, por citar algunos. Todos ellos tienen incorporadas estructuras y operaciones matriciales, fundamentales en el Análisis Multivariante, además de innumerables subrutinas para cálculos más específicos. Puede parecer que el uso de estos programas añade complicaciones a la comprensión de las técnicas expuestas. Pero, en base a la experiencia, hay que decir que ocurre justamente lo contrario: el lenguaje de programación que utilizan se asemeja considerablemente a la notación matricial, lo que contribuye a una mayor asimilación y aprendizaje de las mismas.

Amparo y Aurea

[1]Matlab es una marca registrada de The MathWorks, Inc., http://www.mathworks.com

Contenido

CAPÍTULO 1

Álgebra matricial básica

En este primer capítulo se repasan algunos conceptos de álgebra matricial que serán extremadamente útiles para el tratamiento de datos multivariantes. Las matrices ayudan a plantear los métodos de estadística multivariante de manera concisa y facilitan su implementación en programas de ordenador.
Comenzaremos trabajando con normas de vectores, productos escalares y proyecciones ortogonales. A continuación recordaremos el cálculo de matrices inversas, determinantes, autovalores y autovectores y otros conceptos básicos del álgebra de matrices. El capítulo concluye determinando el signo de algunas formas cuadráticas.

PROBLEMA 1.1

Sean $\mathbf{u} = (1, 2)'$, $\mathbf{v} = (-2, 3)'$ *y* $\mathbf{w} = (3, -5)'$ *tres vectores de* $\mathbb{R}^2$. *Evalúense las siguientes expresiones, donde* $\mathbf{a} \cdot \mathbf{b}$ *denota el producto escalar entre los vectores* $\mathbf{a}$ *y* $\mathbf{b}$ *y* $\|\mathbf{a}\| = \sqrt{\mathbf{a} \cdot \mathbf{a}}$ *denota la norma o longitud del vector* $\mathbf{a}$.

(a) $(\mathbf{u} - 2\mathbf{v}) \cdot \mathbf{w}$

(b) $\|\mathbf{u} + \mathbf{v} + \mathbf{w}\|$

(c) $\|\mathbf{u}\| + \|\mathbf{v}\| + \|\mathbf{w}\|$

(d) $(\mathbf{u} - \mathbf{v}) \cdot (\mathbf{v} - \mathbf{w})$

SOLUCIÓN

Para introducir los vectores en Matlab escribimos

```
u = [1 ; 2];  v = [-2 ; 3];  w = [3 ; -5];
```

(a) $(\mathbf{u} - 2\,\mathbf{v}) \cdot \mathbf{w} = (\mathbf{u} - \mathbf{2}\,\mathbf{v})'\mathbf{w} = \mathbf{35}$. Para calcularlo en Matlab escribimos

```
(u-2*v)'*w
```

(b) $\|\mathbf{u}+\mathbf{v}+\mathbf{w}\| = 2$. Para calcular la norma de un vector $\mathbf{u}$ podremos utilizar la orden de Matlab `norm(u)`. También podemos escribir el código nosotros mismos mediante una función Matlab, que denominaremos, por ejemplo, `norma`. Para utilizar esta función dentro de Matlab, la guardaremos en un fichero con el mismo nombre y extensión `.m`, en este caso `norma.m`:

```
function nu = norma(u)
    u = u(:) ;
    nu = sqrt(u'*u) ;
```

Para resolver este apartado, en la ventana de comandos de Matlab escribiremos:

```
norma(u+v+w)
```

Compruébese que se llega al mismo resultado utilizando la función interna de Matlab `norm`.

(c) $\|\mathbf{u}\| + \|\mathbf{v}\| + \|\mathbf{w}\| = 2.2361$. En Matlab

```
norm(u) + norm(v) + norm(w)
```

(d) $(\mathbf{u}-\mathbf{v})\cdot(\mathbf{v}-\mathbf{w}) = (\mathbf{u}-\mathbf{v})'(\mathbf{v}-\mathbf{w}) = \mathbf{-23}$. Con Matlab se calcularía así

```
(u-v)'*(v-w)
```

PROBLEMA 1.2

Dados dos vectores de $\mathbb{R}^p$, $\mathbf{u}$ y $\mathbf{a}$, encuéntrese la proyección ortogonal del vector $\mathbf{u}$ sobre el vector $\mathbf{a}$, para:

(a) $\mathbf{u} = (8,3)'$, $\quad \mathbf{a} = (4,-5)'$,

(b) $\mathbf{u} = (2,1,-4)'$, $\quad \mathbf{a} = (-5,3,11)'$.

SOLUCIÓN

La proyección ortogonal de $\mathbf{u}$ sobre la dirección determinada por $\mathbf{a}$ viene dada por el vector (Figura 1.1):

$$\mathbf{v} = \frac{\mathbf{u}\cdot\mathbf{a}}{\|\mathbf{a}\|^2}\,\mathbf{a} = (\mathbf{u}\cdot\mathbf{c})\,\mathbf{c},$$

donde $\mathbf{c} = \mathbf{a}/\|\mathbf{a}\|$ es el vector de longitud 1 en la dirección de $\mathbf{a}$. Por tanto, $\mathbf{u}\cdot\mathbf{c}$ es la longitud de la proyección $\mathbf{v}$ (esto lo utilizaremos en el Problema 2.9).
El siguiente código (que debe guardarse en el fichero `ProyOrto.m`) permite calcular la proyección ortogonal de un vector $\mathbf{u}$ sobre $\mathbf{a}$:

```
function v = ProyOrto(u,a)
    u = u(:); a = a(:);
    v = (u'*a)*a /norm(a) ;
```

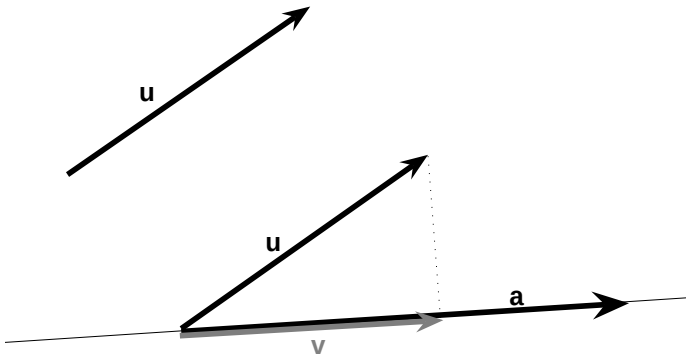


Figura 1.1.
El vector $\mathbf{v}$ es la proyección ortogonal de $\mathbf{u}$ sobre $\mathbf{a}$.

(a) Dentro de Matlab escribimos:

```
u = [8,3]';  a = [4,-5]';
v = ProyOrto(u)
```

y obtenemos $\mathbf{v} = (1.6585, -2.0732)'$.

(b) Análogamente, haciendo:

```
u = [2,1,-4]'; a = [-5,3,11]';
v = ProyOrto(u,a)
```

obtenemos $\mathbf{v} = (1.6452, -0.9871, -3.6194)'$.

PROBLEMA 1.3

Calcúlense los valores de k que hacen que los siguientes vectores $\mathbf{u}$ y $\mathbf{v}$ sean ortogonales.

(a) $\mathbf{u} = (-2, k, -4)', \quad \mathbf{v} = (-1, 3, k)',$

(b) $\mathbf{u} = (-2, k, -k)', \quad \mathbf{v} = (1, 3, k)'.$

SOLUCIÓN

Los vectores $\mathbf{u}$ y $\mathbf{v}$ son ortogonales (o perpendiculares) entre sí, si su producto escalar

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}'\mathbf{v} = \mathbf{v}'\mathbf{u}$$

es 0. Estableciendo esta condición sobre los vectores $\mathbf{u}$ y $\mathbf{v}$ del enunciado, obtendremos una ecuación de la que despejaremos k.

(a) $0 = \mathbf{u}'\mathbf{v} = (-2, k, -4)\begin{pmatrix} -1 \\ 3 \\ k \end{pmatrix} = 2 + 3k - 4k = 2 - k \Rightarrow k = 2.$

(b) $0 = \mathbf{u}'\mathbf{v} = -k^2 + 3k - 2 \Rightarrow k = \dfrac{3 \pm \sqrt{9 - 4(-1)(-2)}}{2} = 2 \;\text{ ó }\; 1.$

PROBLEMA 1.4

Calcúlese la inversa de las matrices

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 4 & 0 \\ \frac{1}{2} & 3 & 2 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 9 & 1 & 0 & 0 \\ 0 & 8 & -2 & 0 \\ 0 & 0 & 7 & -3 \\ 0 & 0 & 0 & 6 \end{pmatrix}.$$

SOLUCIÓN

Uno de los objetivos de este ejercicio es comprobar que la inversa de una matriz triangular inferior (resp. superior) es también una matriz triangular inferior (resp. superior). Recordemos que la inversa de una matriz se calcula mediante la fórmula

$$\mathbf{A}^{-1} = |\mathbf{A}|^{-1} \operatorname{adj}(\mathbf{A}'),$$

donde $|\cdot|$ y $\operatorname{adj}(\cdot)$ denotan, respectivamente, el determinante y la matriz adjunta. Concretamente $|\mathbf{A}| = 8$ y

$$\mathbf{A}^{-1} = \frac{1}{8}\begin{pmatrix} 8 & 0 & 0 \\ -\frac{2}{3} & 2 & 0 \\ -1 & -3 & 4 \end{pmatrix}.$$

Para hacer estos cálculos en Matlab escribimos las siguientes líneas de código

```
A = [ 1   0   0
     1/3  4   0
     1/2  3   2 ] ;
Inv_A = inv(A)
```

El determinante se calcula mediante `det(A)`. Análogamente, $|\mathbf{B}| = 3024$ y

$$\mathbf{B}^{-1} = \frac{1}{1512}\begin{pmatrix} 168 & -21 & -6 & -3 \\ 0 & 189 & 54 & 27 \\ 0 & 0 & 216 & 108 \\ 0 & 0 & 0 & 252 \end{pmatrix}.$$

PROBLEMA 1.5

Considérense las matrices

$$\mathbf{A} = \begin{pmatrix} 4 & 4.001 \\ 4.001 & 4.002 \end{pmatrix} \quad y \quad \mathbf{B} = \begin{pmatrix} 4 & 4.001 \\ 4.001 & 4.002001 \end{pmatrix}.$$

Obsérvese que estas matrices son casi idénticas excepto por una pequeña diferencia en el elemento (2,2). Sin embargo, compruébese que $\mathbf{A}^{-1} \simeq -3\mathbf{B}^{-1}$*, es decir, que pequeños cambios (tal vez debidos al redondeo en las operaciones) pueden dar lugar a inversas muy diferentes.*

SOLUCIÓN

Calculamos las inversas con Matlab

```
A = [ 4 4.001 ; 4.001 4.002 ] ;
Inv_A = inv(A)
B = [ 4 4.001 ; 4.001 4.002001] ;
Inv_B = inv(B)
```

y obtenemos

$$\mathbf{A}^{-1} = 10^6 \begin{pmatrix} -4.0020 & 4.0010 \\ 4.0010 & -4.0000 \end{pmatrix}, \quad \mathbf{B}^{-1} = 10^6 \begin{pmatrix} 1.3340 & -1.3337 \\ -1.3337 & 1.3333 \end{pmatrix}.$$

PROBLEMA 1.6

Calcúlense la ecuación característica y los autovalores de las siguientes matrices

(a) $\mathbf{A}_1 = \begin{pmatrix} 1 & 2 \\ 2 & -2 \end{pmatrix}$,

(b) $\mathbf{A}_2 = \begin{pmatrix} -2 & 0 & 3 \\ 2 & 4 & 0 \\ 1 & 0 & 0 \end{pmatrix}$,

(c) $\mathbf{A}_3 = \begin{pmatrix} 2 & 2 & 2 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$,

(d) $\mathbf{A}_4 = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$.

SOLUCIÓN

(a) Los autovalores de $\mathbf{A}_1$ son las raíces de su polinomio característico

$$P(\lambda) = |\mathbf{A}_1 - \lambda\mathbf{I}| = \begin{vmatrix} 1-\lambda & 2 \\ 2 & -2-\lambda \end{vmatrix} = (1-\lambda)(-2-\lambda) - 4 = \lambda^2 + \lambda - 6.$$

El polinomio $P(\lambda)$ toma el valor 0 para $\lambda_1 = 2$ ó $\lambda_2 = -3$. Éstos son los autovalores de $\mathbf{A}_1$ (conviene ya acostumbrarse a ordenarlos de mayor a menor, pues más adelante, al calcular componentes principales, será necesario). La ecuación característica es la que se obtiene de igualar el polinomio característico a cero $P(\lambda) = 0$, es decir, la ecuación $\lambda^2 + \lambda - 6 = 0$.

(b) El polinomio característico de $\mathbf{A}_2$ es

$$\begin{aligned} P(\lambda) &= |\mathbf{A}_2 - \lambda\mathbf{I}| \\ &= \begin{vmatrix} -2-\lambda & 0 & 3 \\ 2 & 4-\lambda & 0 \\ 1 & 0 & -\lambda \end{vmatrix} \\ &= (\lambda-4)(3-2\lambda-\lambda^2) = (\lambda-4)(\lambda+3)(1-\lambda). \end{aligned}$$

Por tanto, la ecuación característica de $\mathbf{A}_2$ es $(\lambda-4)(\lambda+3)(1-\lambda) = 0$. Los autovalores de $\mathbf{A}_2$ son la solución de la ecuación anterior: $\lambda_1 = 4$, $\lambda_2 = 1$ y $\lambda_3 = -3$.

(c) La ecuación característica de $\mathbf{A}_3$ es $0 = P(\lambda) = |\mathbf{A}_3 - \lambda\mathbf{I}| = \lambda^2(4-\lambda)$. Entonces los autovalores de $\mathbf{A}_3$ son $\lambda_1 = 4$ (autovalor simple) y $\lambda_2 = \lambda_3 = 0$ (autovalor doble).

(d) La ecuación característica de $\mathbf{A}_4$ es $0 = (\lambda-1)^2(\lambda-4)$, por lo que sus autovalores son $\lambda_1 = 4$ y $\lambda_2 = \lambda_3 = 1$ (autovalor doble).

PROBLEMA 1.7

Genérese una matriz $\mathbf{X}$*, de dimensión* 4×3 *y un vector* $\mathbf{u}$*,* 4×1*, ambos de números aleatorios y constrúyanse las matrices simétricas* $\mathbf{A} = \mathbf{X}'\,\mathbf{X}$ *y* $\mathbf{B} = \mathbf{u}\,\mathbf{u}'$.

(a) *Calcúlense la traza y el determinante de* $\mathbf{A}$ *y* $\mathbf{B}$.

(b) *Obténganse los autovalores y autovectores de* $\mathbf{A}$ *y* $\mathbf{B}$.

(c) *Compruébese que la traza y el determinante de* $\mathbf{A}$ *coinciden respectivamente con la suma y el producto de los autovalores de* $\mathbf{A}$.

(d) *Obténganse los rangos de* $\mathbf{A}$ *y* $\mathbf{B}$ *y compruébese que coinciden, respectivamente, con el número de autovalores no nulos de* $\mathbf{A}$ *y* $\mathbf{B}$.

SOLUCIÓN

Empezamos construyendo las matrices $\mathbf{A}$ y $\mathbf{B}$ a partir de la generación aleatoria de $\mathbf{X}$ y $\mathbf{u}$:

```
X = rand[4,3];
u = rand[4,1];
A = X'*X;
B = u*u';
```

(a) Las instrucciones `trace(A)` y `det(A)` permiten obtener la traza y el determinante de $\mathbf{A}$. Haremos lo mismo para $\mathbf{B}$.

(b) La instrucción `[T,D]=eig(A)` permite encontrar la descomposición espectral de $\mathbf{A}$, es decir, $\mathbf{A} = \mathbf{T}\,\mathbf{D}\,\mathbf{T}'$, donde $\mathbf{D}$ y $\mathbf{T}$ son matrices de la misma dimensión que $\mathbf{A}$, tales que: $\mathbf{D}$ es una matriz diagonal que contiene los autovalores de $\mathbf{A}$, y $\mathbf{T}$ es una matriz ortogonal (es decir, $\mathbf{T}\,\mathbf{T}' = \mathbf{T}'\,\mathbf{T} = \mathbf{I}$) cuyas columnas son los autovectores de $\mathbf{A}$.
Utilizando la misma instrucción obtendremos los autovalores y autovectores de $\mathbf{B}$. Observad que la matriz diagonal que contiene los autovalores de $\mathbf{B}$ tiene solamente un elemento diagonal no nulo.

(c) Hay que comprobar que la suma y el producto de la diagonal de la matriz $\mathbf{D}$, es decir, `sum(diag(D))` y `prod(diag(D))`, coinciden con `trace(A)` y `det(A)`, respectivamente.

(d) La instrucción `rank(A)` permite obtener el rango de $\mathbf{A}$, que debe coincidir con el número de elementos no nulos de la diagonal de $\mathbf{D}$. Haremos lo mismo para $\mathbf{B}$. Observad que $\mathbf{B}$ es una matriz de rango uno, tal como cabía esperar, puesto que la hemos construido a partir de un único vector.

PROBLEMA 1.8

Considérense las matrices siguientes:

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 4 \\ -1 & 4 & 1 \\ 2 & -1 & 4 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}.$$

(a) *¿Son idempotentes?*

(b) *Calcúlese su determinante.*

(c) *¿Son definidas positivas?*

(d) *¿Son ortogonales?*

SOLUCIÓN

(a) Una matriz cuadrada $\mathbf{A}$ es idempotente si $\mathbf{A}^2 = \mathbf{A}$. En este caso, o bien $\mathbf{A}$ es la matriz identidad, o bien $\mathbf{A}$ es singular (es decir, $|\mathbf{A}| = 0$). Asímismo, si $\mathbf{A}$ es idempotente entonces $\text{rg}(\mathbf{A}) = \text{tr}(\mathbf{A})$.
Puesto que $|\mathbf{A}| = 12 \neq 0$ y $|\mathbf{C}| = 6 \neq 0$, entonces ni $\mathbf{A}$, ni $\mathbf{C}$ son idempotentes. Por otro lado, aunque $|\mathbf{B}| = 0$, la matriz $\mathbf{B}$ tampoco es idempotente, porque $tr(\mathbf{B}) = 3 \neq rg(\mathbf{B}) = 2$.

(b) Está respondido en el apartado anterior.

(c) Los menores principales de $\mathbf{A}$ son

$$\begin{aligned} |2| &= 2 > 0\,; \\ \left|\begin{array}{rr} 2 & 1 \\ -1 & 4 \end{array}\right| &= 9 > 0\,; \\ |\mathbf{A}| &= 12 > 0. \end{aligned}$$

Por tanto, por el criterio de Sylvester, $\mathbf{A}$ es definida positiva. En cambio, $\mathbf{B}$ no lo es puesto que $|\mathbf{B}| = 0$. Para ver que $\mathbf{C}$ es definida positiva podemos calcular sus autovalores con Matlab:

```
C = [ 2 1 1 ; 1 2 -1 ; -1 -1 2 ];
eig(C)
```

y vemos que todos son positivos $\lambda_1 = 3$, $\lambda_2 = 2$ y $\lambda_3 = 1$. Por tanto, $\mathbf{C}$ es definida positiva.

(d) Una matriz cuadrada $\mathbf{A}$ es ortogonal si

$$\mathbf{A}\,\mathbf{A}' = \mathbf{A}'\,\mathbf{A} = \mathbf{I}\,.$$

Con el código `A*A'`, `B*B'`, `C*C'`, comprobamos que ninguna de las tres matrices verifica esta condición y, por tanto, ni $\mathbf{A}$, ni $\mathbf{B}$, ni $\mathbf{C}$ son ortogonales. Por ejemplo,

$$\mathbf{A}\,\mathbf{A}' = \begin{pmatrix} 21 & 6 & 19 \\ 6 & 18 & -2 \\ 19 & -2 & 21 \end{pmatrix}.$$

PROBLEMA 1.9

Calcúlese la descomposición espectral de

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{pmatrix}.$$

SOLUCIÓN

La descomposición espectral de una matriz simétrica $\mathbf{A}$ de dimensión $k \times k$ consiste en expresar $\mathbf{A}$ de la siguiente manera:

$$\mathbf{A} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1' + \lambda_2 \mathbf{e}_2 \mathbf{e}_2' + \ldots + \lambda_k \mathbf{e}_k \mathbf{e}_k', \tag{1.1}$$

donde $\lambda_1, \ldots, \lambda_k$ son los autovalores de $\mathbf{A}$ y $\mathbf{e}_1, \ldots, \mathbf{e}_k$ son autovectores normalizados de $\mathbf{A}$ asociados respectivamente a $\lambda_1, \ldots, \lambda_k$ y ortogonales entre sí. Recordemos que esta última condición se cumple automáticamente en una matriz simétrica cuando todos sus autovalores son distintos. Sin embargo, cuando hay algún autovalor múltiple (como en este caso) hay que escoger los autovectores adecuadamente.

Los autovalores de $\mathbf{A}$ son las raíces de la ecuación característica

$$0 = |\mathbf{A} - \lambda\mathbf{I}| = (\lambda - 1)^2(7 - \lambda),$$

es decir, son $\lambda_1 = 7$ y $\lambda_2 = \lambda_3 = 1$. Un autovector $\mathbf{x}$ de $\mathbf{A}$ asociado al autovalor λ es un vector que verifica la ecuación

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}\,.$$

Por ejemplo, para $\lambda_1 = 7$, buscamos un vector $\mathbf{x} = (x_1, x_2, x_3)'$ tal que

$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \left(\begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{pmatrix} - 7 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$
$$= \begin{pmatrix} -4 & 2 & 2 \\ 2 & -4 & 2 \\ 2 & 2 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

lo cual equivale al sistema de ecuaciones

$$\begin{aligned} 0 &= -2x_1 + x_2 + x_3, \\ 0 &= x_1 - 2x_2 + x_3. \end{aligned}$$

De este sistema deducimos que un autovector $\mathbf{x}$ correspondiente al autovalor $\lambda_1 = 7$ debe cumplir la condición $x_1 = x_2 = x_3$. Por ejemplo, podríamos tomar el vector $(1, 1, 1)'$. Un autovector normalizado de $\mathbf{A}$ correspondiente al autovalor $\lambda_1 = 8$ es, pues, $\mathbf{e}_1 = (1, 1, 1)'/\sqrt{3}$. Respecto al autovalor $\lambda_2 = 1$, la ecuación

$$(\mathbf{A} - \lambda_2\mathbf{I})\,\mathbf{x} = \mathbf{0}$$

implica

$$x_1 + x_2 + x_3 = 0. \tag{1.2}$$

Observemos que el número de condiciones que debe cumplir un autovector de $\mathbf{A}$ es $\mathrm{rg}(\mathbf{A})$, el rango de $\mathbf{A}$, menos la multiplicidad del autovalor correspondiente. En este caso hay sólo una ecuación, pues $\mathrm{rg}(\mathbf{A}) = 3$ y $\lambda = 1$ es un autovalor doble. Para la descomposición espectral es necesario que todos los autovectores $\mathbf{e}_i$ sean ortogonales entre sí, luego debemos buscar dos vectores que verifiquen la condición (1.2) y cuyo producto escalar sea cero. Por ejemplo, $\mathbf{e}_2 = (1, -1, 0)'/\sqrt{2}$ y $\mathbf{e}_3 = (1, 1, -2)'/\sqrt{6}$.
Así pues la descomposición espectral de la matriz $\mathbf{A}$ es:

$$\mathbf{A} = \frac{1}{2} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} (1, -1, 0) + \frac{1}{6} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} (1, 1, -2) + \frac{7}{3} \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} (1, 1, 2).$$

Observación. La definición 1.1 admite una expresión en forma matricial, tal y como vimos en el Problema 1.7. Dejamos al lector que escriba la descomposición espectral de $\mathbf{A}$ como un producto de 3 matrices cuadradas.

PROBLEMA 1.10

Dada la matriz

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

(a) *Calcúlense sus autovalores, los de* $\mathbf{A}^2$ *y los de* $\mathbf{A}^{-1}$.

(b) *Calcúlese una base ortogonal que la diagonalice.*

SOLUCIÓN

Puesto que $\mathbf{A}$ es una matriz simétrica, el teorema de descomposición espectral asegura que existen una matriz ortogonal $\mathbf{T}$ y una matriz diagonal $\mathbf{\Lambda}$ tales que $\mathbf{A} = \mathbf{T}\,\mathbf{\Lambda}\,\mathbf{T}'$. La matriz $\mathbf{\Lambda}$ contiene los autovalores de $\mathbf{A}$ y la matriz $\mathbf{T}$ contiene los autovectores de $\mathbf{A}$. Además se verifica la siguiente propiedad:

$$\mathbf{A}^p = \mathbf{T}\,\mathbf{\Lambda}^p\,\mathbf{T}',$$

para $p \in \mathbb{Z}$.
Mediante Matlab, obtenemos la descomposición espectral de $\mathbf{A}$ y comprobamos la propiedad anterior para $p = 2$ y $p = -1$

```
A = [3 2 0; 2 3 0; 0 0 3];
[T,Lambda] = eig(A);
```

Los resultados que se obtienen son:

$$\mathbf{T} = \begin{pmatrix} -0.7071 & 0 & 0.7071 \\ 0.7071 & 0 & 0.7071 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{\Lambda} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

Observad que las columnas de `T` forman una base ortogonal que diagonaliza a la matriz $\mathbf{A}$. Calculamos los autovalores de $\mathbf{A}^2$ y de $\mathbf{A}^{-1}$ con:

```
Lambda2 = eig(A*A);
Lambdainv = eig(inv(A));
```

y obtenemos que los autovalores de $\mathbf{A}^2$ son 1, 9 y 25 y los de $\mathbf{A}^{-1}$ son 1, 0.33 y 0.2. Podéis comprobar que las instrucciones:

```
T*diag(Lambda2)*T'
T*diag(Lambdainv)*T'
```

permiten recuperar las matrices $\mathbf{A}^2$ y $\mathbf{A}^{-1}$ respectivamente.

PROBLEMA 1.11

Considérese la matriz

$$\mathbf{A} = \begin{pmatrix} 2 & a \\ a & 2 \end{pmatrix}.$$

(a) *Calcúlense los autovalores y autovectores de* $\mathbf{A}$.

(b) *¿Para qué valores de* a *es la matriz* $\mathbf{A}$ *definida positiva?*

SOLUCIÓN

(a) Los autovalores de $\mathbf{A}$ son $\lambda_1 = 2+|a|$ y $\lambda_2 = 2-|a|$. Los correspondientes autovectores normalizados son $\mathbf{e}_1 = (\text{sgn}(a), 1)'/\sqrt{2}$ y $\mathbf{e}_2 = (1, -\text{sgn}(a))'/\sqrt{2}$, siendo $\text{sgn}(a) = a/|a|$ el signo de a.

(b) $\mathbf{A}$ es definida positiva si y sólo si sus autovalores son ambos positivos, es decir, si $|a| < 2$.

PROBLEMA 1.12

Considérese la siguiente matriz

$$\mathbf{A} = \begin{pmatrix} 6 & 10 \\ 10 & 6 \\ 1 & 5 \end{pmatrix}.$$

(a) *Encuéntrese la inversa generalizada de Moore-Penrose,* $\mathbf{A}^-$, *de* $\mathbf{A}$.

(b) *Compruébese que se cumple la propiedad*

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}. \tag{1.3}$$

(c) *Compruébese que se cumplen las propiedades*

(i) $\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-$,

(ii) $\mathbf{A}^-\mathbf{A}$ *es simétrica,*

(iii) $\mathbf{A}\mathbf{A}^-$ *es simétrica.*

SOLUCIÓN

(a) La inversa de Moore-Penrose es aquella matriz $\mathbf{A}^-$ que verifica las condiciones (1.3) y (i)–(iii) del apartado (c). La matriz $\mathbf{A}^-$ se obtiene a partir de la descomposición en valores singulares de

$$\mathbf{A} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{V}',$$

es decir,

$$\mathbf{A}^{-} = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{U}'.$$

La función Matlab que calcule esta inversa podría ser

```
function B = ginvMP(A)
   [U,D,V] = svd(A,0) ;
   B = V*inv(D)*U' ;
```

También podemos utilizar directamente la función `B = pinv(A)` implementada ya en Matlab. El resultado es:

$$\mathbf{A}^{-} = \begin{pmatrix} -0.0442 & 0.1337 & -0.0721 \\ 0.0964 & -0.0665 & 0.0871 \end{pmatrix}.$$

(b) La expresión (1.3) es la propiedad que tiene que cumplir cualquier inversa generalizada. Para comprobar con Matlab que se cumple escribimos:

```
B = ginvMP(A) ;
A*B*A
```

(c) Las propiedades (i)–(iii) del apartado (c) se comprueban escribiendo las instrucciones `B*A*B`, `B*A` y `A*B`. El primer producto proporciona la matriz $\mathbf{B}$ y el segundo y tercero dan, respectivamente:

$$\mathbf{B}\,\mathbf{A} = \mathbf{I}, \quad \mathbf{A}\,\mathbf{B} = \begin{pmatrix} 0.6990 & 0.1368 & 0.4378 \\ 0.1368 & 0.9378 & -0.1990 \\ 0.4378 & -0.1990 & 0.3632 \end{pmatrix},$$

que son matrices simétricas, donde $\mathbf{I}$ es la matriz identidad 2×2.

PROBLEMA 1.13

Calcúlese la matriz simétrica asociada a cada una de las siguientes formas cuadráticas y determínese si es definida positiva.

(a) $Q(x_1, x_2) = 2x_1^2 - 3x_1x_2 + 3x_2^2$,

(b) $Q(x_1, x_2, x_3) = x_1^2 + x_1x_3 + 0.25x_3^2 + 1.6x_1x_2 + 0.6x_2^2 + 0.8x_2x_3$.

SOLUCIÓN

(a) La matriz simétrica

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$$

asociada a Q es la que verifica $Q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$, donde $\mathbf{x} = (x_1, x_2)'$. Como

$$\mathbf{x}'\mathbf{A}\mathbf{x} = (x_1, x_2)\mathbf{A}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2,$$

tenemos que $a_{11} = 2$, $a_{22} = 3$, $2\,a_{12} = -3$. Por tanto,

$$\mathbf{A} = \begin{pmatrix} 2 & -3/2 \\ -3/2 & 3 \end{pmatrix}.$$

Para comprobar que $\mathbf{A}$ es definida positiva, en Matlab escribimos:

```
A = [2 -3/2 ; -3/2 3] ;
lambda = eig(A)'
```

que nos proporciona los autovalores 0.9189 y 4.0811, ambos positivos.

(b) La matriz simétrica

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}$$

asociada a Q es la que verifica $Q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$, donde $\mathbf{x} = (x_1, x_2, x_3)'$. Como

$$\mathbf{x}'\,\mathbf{A}\,\mathbf{x} = a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + 2a_{23}x_2x_3,$$

tenemos que:

$$\mathbf{A} = \begin{pmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 0.6 & 0.4 \\ 0.5 & 0.4 & 0.25 \end{pmatrix}.$$

Calculando los autovalores de $\mathbf{A}$ obtenemos que uno de ellos es negativo, -0.0266, por lo que $\mathbf{A}$ no es definida positiva.

PROBLEMA 1.14

Sean $\mathbf{x} = (x_1, x_2)'$ un vector y $Q(\mathbf{x}) = m\,x_2^2 - 4\,x_1\,x_2 + x_1^2$ una forma cuadrática, donde $m \in \mathbb{R}$.

(a) *Determínese la matriz simétrica $\mathbf{A}$ asociada a $Q(\mathbf{x})$.*

(b) *Determínense los valores de m para que $\mathbf{A}$ sea definida positiva.*

(c) *Hállense los autovalores y los autovectores asociados a $\mathbf{A}$ en el caso de que $m = -2$.*

SOLUCIÓN

(a) $\mathbf{A} = \begin{pmatrix} 1 & -2 \\ -2 & m \end{pmatrix}$.

(b) $\mathbf{A}$ es definida positiva si y sólo si todos los menores principales tienen determinante positivo. Por tanto, $m > 4$.

(c) Para el caso $m = -2$, los autovalores de $\mathbf{A}$ son $\lambda_1 = 2$ y $\lambda_2 = -3$. Los autovectores normalizados son respectivamente $\mathbf{e}_1 = (-2, 1)'/\sqrt{5}$ y $\mathbf{e}_2 = (1, 2)'/\sqrt{5}$.

PROBLEMA 1.15

Considérense las siguientes matrices simétricas de dimensión 3×3*:*

$$\mathbf{A} = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix} \quad y \quad \mathbf{B} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

(a) *Decídase el signo de la forma cuadrática* $q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$*, donde* $\mathbf{x} \in \mathbb{R}^3$.

(b) *Escríbase la expresión explícita de la forma cuadrática* $Q(\mathbf{x}) = q(\mathbf{x}) + \mathbf{x}'\mathbf{B}\mathbf{x}$. *Sin calcular los autovalores de* $\mathbf{A} + \mathbf{B}$ *decídase el signo de* $Q(\mathbf{x})$.

SOLUCIÓN

(a) Con el mismo código que utilizamos en el Problema 1.13 podemos ver que los autovalores de $\mathbf{A}$ son $\lambda_1 = 4$, $\lambda_2 = 3$ y $\lambda_3 = 2$. Por tanto, $\mathbf{A}$ y su forma cuadrática, q, son definidas positivas.

(b) La forma Q es definida positiva porque q lo es y $\mathbf{B}$ es semidefinida positiva. Es decir, puesto que $q(\mathbf{x}) > 0$ y $\mathbf{x}'\mathbf{B}\mathbf{x} \geq 0$ para $\mathbf{x} \neq \mathbf{0}$, entonces se verifica que $Q(\mathbf{x}) > 0$ para $\mathbf{x} \neq 0$.

CAPÍTULO 2

Estadísticos descriptivos

Los objetivos de este capítulo son sencillos, pero fundamentales (en cuanto a notación y conceptos) para la posterior comprensión de los capítulos restantes. Aprenderemos a manejar datos multivariantes de manera matricial y a representarlos gráficamente. Calcularemos las medidas resumen más utilizadas de localización, dispersión y dependencia muestrales: el vector de medias, la matriz de varianzas-covarianzas y la matriz de correlaciones. A lo largo del tema se insiste en la interpretación intuitiva de estos estadísticos y de los gráficos. Quedará patente la utilidad de Matlab para el tratamiento de datos multidimensionales. También se hace especial hincapié en el cálculo de combinaciones lineales de los vectores observados.

PROBLEMA 2.1

Se define la matriz de centrado *de dimensión* n *como* $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\,\mathbf{1}'$, *donde* $\mathbf{I}$ *es la matriz identidad de dimensión* $n \times n$ *y* $\mathbf{1}$ *es un vector* $n \times 1$ *de unos. La utilidad de esta matriz* $\mathbf{H}$ *radica en que, como su nombre indica, se usa para centrar configuraciones de datos: si* $\mathbf{X}$ *es una matriz de datos de dimensión* $n \times p$, *entonces* $\mathbf{H}\,\mathbf{X}$ *es una matriz cuyas columnas tienen media cero.*
Utilícese Matlab para comprobar las dos siguientes propiedades de la matriz de centrado (tomando, por ejemplo, $n = 5$*):*

(a) $\mathbf{H}$ *es idempotente,.*

(b) *rg*$(\mathbf{H}) =$ *tr*$(\mathbf{H}) = n - 1$.

SOLUCIÓN

Construimos la matriz de centrado de dimensión $n = 5$:

```
n = 5;
H = eye(n)-ones(n,n)/n;
```

y comprobamos que `H^ 2` coincide con `H`. Las instrucciones `trace(H)` y `rank(H)` permiten obtener su traza y su rango, que deben ser $n - 1 = 4$.

PROBLEMA 2.2

Los datos de la Tabla 2.1 corresponden a chalets construidos por diez promotoras que operan a lo largo de la costa española.

Tabla 2.1.
Diez promotoras de la costa española (Problema 2.2)

Promotora	X_1 =Duración media hipoteca (años)	X_2 =Precio medio (millones euros)	X_3 =Superficie media (m^2) de cocina
1	8.7	0.3	3.1
2	14.3	0.9	7.4
3	18.9	1.8	9.0
4	19.0	0.8	9.4
5	20.5	0.9	8.3
6	14.7	1.1	7.6
7	18.8	2.5	12.6
8	37.3	2.7	18.1
9	12.6	1.3	5.9
10	25.7	3.4	15.9

(a) *Dibújese el diagrama de dispersión múltiple y coméntese el aspecto del gráfico.*

(b) *Para X_1 y X_2 calcúlense, respectivamente, las medias muestrales $\bar{x}_1$ y $\bar{x}_2$, las varianzas muestrales s_{11} y s_{22}, la covarianza entre X_1 y X_2, s_{12}, y la correlación entre ambas, r_{12}. Interprétese el valor obtenido de r_{12}.*

(c) *Utilizando la matriz de datos $\mathbf{X}$ y la de centrado $\mathbf{H}$ definida en el Problema 2.1, calcúlense el vector de medias muestrales $\bar{\mathbf{x}}$ y la matriz de covarianzas muestrales $\mathbf{S}$. A partir de ésta obténgase la matriz de correlaciones $\mathbf{R}$.*

SOLUCIÓN

(a) En la Figura 2.1 se puede ver el diagrama de dispersión múltiple de las tres variables. Se observa que todas ellas están positivamente correladas entre sí y que el grado de correlación es muy alto. Por tanto, una sola de esas variables debería poder servir para predecir cualquiera de las otras dos.
Las instrucciones en Matlab para introducir los datos y realizar el gráfico son

```
X = [ 8.7    0.3    3.1
      14.3   0.9    7.4
      18.9   1.8    9.0
      19.0   0.8    9.4
      20.5   0.9    8.3
      14.7   1.1    7.6
      18.8   2.5    12.6
      37.3   2.7    18.1
      12.6   1.3    5.9
      25.7   3.4    15.9];

plotmatrix(X)
```

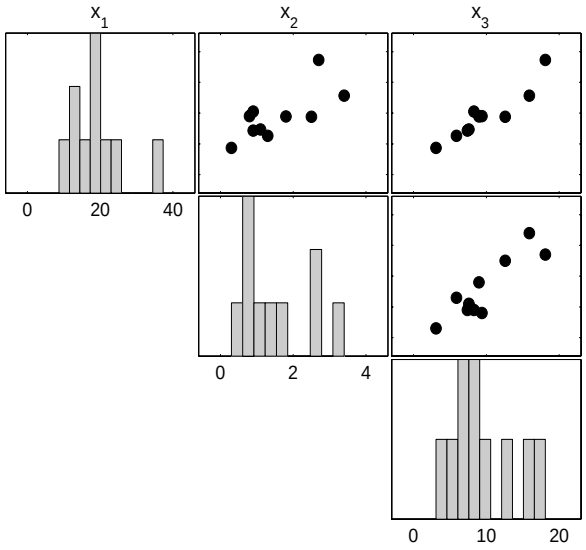


Figura 2.1.
Datos de chalets construidos por promotoras (Problema 2.2)

(b) Para calcular con Matlab los valores de

$$\bar{x}_1 = \frac{1}{10}\sum_{i=1}^{10} x_{i1} = 19.05 \qquad \text{y} \qquad \bar{x}_2 = \frac{1}{10}\sum_{i=1}^{10} x_{i2} = 1.57$$

escribimos el siguiente código:

```
[n,p] = size(X) ;
m1 = sum(X(:,1))/n ;
m2 = sum(X(:,2))/n ;
```

o también

```
m1 = mean(X(:,1)) ;    m2 = mean(X(:,2)) ;
```

Las varianzas

$$s_{11} = \frac{1}{10}\sum_{i=1}^{10} x_{i1}^2 - \bar{x}_1^2 = 56.97 \quad \text{y} \quad s_{22} = \frac{1}{10}\sum_{i=1}^{10} x_{i2}^2 - \bar{x}_2^2 = 0.89$$

se calculan con

```
s11 = sum(X(:,1).^2)/n - m1^2; s22 = sum(X(:,2).^2)/n - m2^2;
```

o bien con

```
s11 = var(X(:,1),1) ;  s22 = var(X(:,2),1) ;
```

Por último, con las instrucciones

```
s12 = sum(X(:,1).*X(:,2))/n - m1*m2 ;
r12 = s12/sqrt(s11*s22) ;
```

obtenemos

$$s_{12} = \frac{1}{10}\sum_{i=1}^{10} x_{i1}x_{i2} - \bar{x}_1\bar{x}_2 = 5.17 \qquad \text{y} \qquad r_{12} = \frac{s_{12}}{\sqrt{s_{11}s_{22}}} = 0.72.$$

El valor de la correlación entre las variables X_1 y X_2 es positivo y alto, como ya permitía deducir el diagrama de dispersión del apartado (a).

(c) Los valores que acabamos de calcular en el apartado (b) para medias, varianzas, covarianzas y correlaciones se pueden obtener matricialmente. La instrucción de Matlab que calcula $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}_n = (19.32, 1.51, 9.76)'$ es:

```
m = X' * ones(n,1)/n ;
```

Para comprobar que

$$\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{H}\mathbf{X} = \begin{pmatrix} 56.97 & 5.17 & 30.48 \\ & 0.89 & 3.65 \\ & & 18.76 \end{pmatrix}$$

escribiremos:

```
H = eye(n)-ones(n,n)/n ;
S = X'*H*X/n ;
```

Por último, la matriz

$$\mathbf{R} = \mathbf{D}^{-1/2}\begin{pmatrix} 1 & 0.71 & 0.95 \\ & 1 & 0.85 \\ & & 1 \end{pmatrix}\mathbf{D}^{-1/2},$$

donde $\mathbf{D}^{-1/2} = \text{diag}(s_{11}^{-1/2}, s_{22}^{-1/2}, s_{33}^{-1/2})$, se obtiene mediante:

```
d = diag(S).^(-0.5) ;
R = diag(d) * S * diag(d) ;
```

Podéis comprobar que las funciones internas de Matlab:

```
m = mean(X) ;      S = cov(X,1) ;      R = corrcoef(X)
```

producen los mismos resultados. Si, en cambio, escribimos `cov(X)` Matlab calcula la matriz de dispersión $\tilde{\mathbf{S}} = \frac{1}{n-1}\mathbf{X}'\mathbf{H}\mathbf{X}$, que a veces se denomina matriz de varianzas-covarianzas corregida.

PROBLEMA 2.3

La contaminación por mercurio de peces de agua dulce comestibles es una amenaza directa contra nuestra salud. Entre 1990 y 1991 se llevó a cabo un estudio en 53 lagos de Florida con el fin de examinar los factores que influían en el nivel de contaminación por mercurio. Las variables que se midieron fueron:

X_1 = *número de identificación,*
X_2 = *nombre del lago,*
X_3 = *alcalinidad (mg/l de carbonato de calcio),*
X_4 = *pH,*
X_5 = *calcio (mg/l),*
X_6 = *clorofila (mg/l),*
X_7 = *concentración media de mercurio (partes por millón) en el tejido músculo del grupo de peces estudiados en cada lago,*
X_8 = *número de peces estudiados por lago,*
X_9 = *mínimo de la concentración de mercurio en cada grupo de peces,*
X_{10} = *máximo de la concentración de mercurio en cada grupo de peces,*
X_{11} = *estimación (mediante regresión) de la concentración de mercurio en un pez de 3 años (o promedio de mercurio cuando la edad no está disponible),*
X_{12} = *indicador de la edad de los peces.*

La Tabla 2.2 contiene los datos de este estudio, disponible en la página web

`http://lib.stat.cmu.edu/DASL.`

(a) *Represéntense de forma conjunta las variables X_3, X_6, X_7 y véase cómo se modifica su dispersión cuando se producen transformaciones (lineales y no lineales) sobre las variables. Considérense como medidas de dispersión global la traza y el determinante de la matriz de covarianzas .*

(b) *Dibújese el histograma tridimensional correspondiente a X_3 y X_7. Elíjanse sendas transformaciones no lineales para estas variables de entre las utilizadas en el apartado anterior y dibújese el histograma tridimensional de las variables transformadas.*

SOLUCIÓN

(a) Supongamos que tenemos un fichero de texto, de nombre `mercurio.txt` que contiene los datos de la Tabla 2.2, cuya primera fila contiene los nombres de las variables, de manera que

Tabla 2.2.

Datos del ejercicio 2.3 (http://lib.stat.cmu.edu/DASL/Datafiles/MercuryinBass.html)

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
1	Alligator	5.9	6.1	3.0	0.7	1.23	5	0.85	1.43	1.53	1
2	Annie	3.5	5.1	1.9	3.2	1.33	7	0.92	1.90	1.33	0
3	Apopka	116.0	9.1	44.1	128.3	0.04	6	0.04	0.06	0.04	0
4	Blue Cypress	39.4	6.9	16.4	3.5	0.44	12	0.13	0.84	0.44	0
5	Brick	2.5	4.6	2.9	1.8	1.20	12	0.69	1.50	1.33	1
6	Bryant	19.6	7.3	4.5	44.1	0.27	14	0.04	0.48	0.25	1
7	Cherry	5.2	5.4	2.8	3.4	0.48	10	0.30	0.72	0.45	1
8	Crescent	71.4	8.1	55.2	33.7	0.19	12	0.08	0.38	0.16	1
9	Deer Point	26.4	5.8	9.2	1.6	0.83	24	0.26	1.40	0.72	1
10	Dias	4.8	6.4	4.6	22.5	0.81	12	0.41	1.47	0.81	1
11	Dorr	6.6	5.4	2.7	14.9	0.71	12	0.52	0.86	0.71	1
12	Down	16.5	7.2	13.8	4.0	0.50	12	0.10	0.73	0.51	1
13	Eaton	25.4	7.2	25.2	11.6	0.49	7	0.26	1.01	0.54	1
14	East Tohopekaliga	7.1	5.8	5.2	5.8	1.16	43	0.50	2.03	1.00	1
15	Farm-13	128.0	7.6	86.5	71.1	0.05	11	0.04	0.11	0.05	0
16	George	83.7	8.2	66.5	78.6	0.15	10	0.12	0.18	0.15	1
17	Griffin	108.5	8.7	35.6	80.1	0.19	40	0.07	0.43	0.19	1
18	Harney	61.3	7.8	57.4	13.9	0.77	6	0.32	1.50	0.49	1
19	Hart	6.4	5.8	4.0	4.6	1.08	10	0.64	1.33	1.02	1
20	Hatchineha	31.0	6.7	15.0	17.0	0.98	6	0.67	1.44	0.70	1
21	Iamonia	7.5	4.4	2.0	9.6	0.63	12	0.33	0.93	0.45	1
22	Istokpoga	17.3	6.7	10.7	9.5	0.56	12	0.37	0.94	0.59	1
23	Jackson	12.6	6.1	3.7	21.0	0.41	12	0.25	0.61	0.41	0
24	Josephine	7.0	6.9	6.3	32.1	0.73	12	0.33	2.04	0.81	1
25	Kingsley	10.5	5.5	6.3	1.6	0.34	10	0.25	0.62	0.42	1
26	Kissimmee	30.0	6.9	13.9	21.5	0.59	36	0.23	1.12	0.53	1
27	Lochloosa	55.4	7.3	15.9	24.7	0.34	10	0.17	0.52	0.31	1
28	Louisa	3.9	4.5	3.3	7.0	0.84	8	0.59	1.38	0.87	1
29	Miccasukee	5.5	4.8	1.7	14.8	0.50	11	0.31	0.84	0.50	0
30	Minneola	6.3	5.8	3.3	0.7	0.34	10	0.19	0.69	0.47	1
31	Monroe	67.0	7.8	58.6	43.8	0.28	10	0.16	0.59	0.25	1
32	Newmans	28.8	7.4	10.2	32.7	0.34	10	0.16	0.65	0.41	1
33	Ocean Pond	5.8	3.6	1.6	3.2	0.87	12	0.31	1.90	0.87	0
34	Ocheese Pond	4.5	4.4	1.1	3.2	0.56	13	0.25	1.02	0.56	0
35	Okeechobee	119.1	7.9	38.4	16.1	0.17	12	0.07	0.30	0.16	1
36	Orange	25.4	7.1	8.8	45.2	0.18	13	0.09	0.29	0.16	1
37	Panasoffkee	106.5	6.8	90.7	16.5	0.19	13	0.05	0.37	0.23	1
38	Parker	53.0	8.4	45.6	152.4	0.04	4	0.04	0.06	0.04	0
39	Placid	8.5	7.0	2.5	12.8	0.49	12	0.31	0.63	0.56	1
40	Puzzle	87.6	7.5	85.5	20.1	1.10	10	0.79	1.41	0.89	1
41	Rodman	114.0	7.0	72.6	6.4	0.16	14	0.04	0.26	0.18	1
42	Rousseau	97.5	6.8	45.5	6.2	0.10	12	0.05	0.26	0.19	1
43	Sampson	11.8	5.9	24.2	1.6	0.48	10	0.27	1.05	0.44	1
44	Shipp	66.5	8.3	26.0	68.2	0.21	12	0.05	0.48	0.16	1
45	Talquin	16.0	6.7	41.2	24.1	0.86	12	0.36	1.40	0.67	1
46	Tarpon	5.0	6.2	23.6	9.6	0.52	12	0.31	0.95	0.55	1
47	Trafford	81.5	8.9	20.5	9.6	0.27	6	0.04	0.40	0.27	0
48	Trout	1.2	4.3	2.1	6.4	0.94	10	0.59	1.24	0.98	1
49	Tsala Apopka	34.0	7.0	13.1	4.6	0.40	12	0.08	0.90	0.31	1
50	Weir	15.5	6.9	5.2	16.5	0.43	11	0.23	0.69	0.43	1
51	Tohopekaliga	25.6	6.2	12.6	27.7	0.65	44	0.30	1.10	0.58	1
52	Wildcat	17.3	5.2	3.0	2.6	0.25	12	0.15	0.40	0.28	1
53	Yale	71.8	7.9	20.5	8.8	0.27	12	0.15	0.51	0.25	1

los datos propiamente dichos empiezan en la segunda fila, escritos por columnas y separados uno de otro mediante tabulación. Para leer los datos desde Matlab utilizaremos la función `dlmread`. Observemos que las dos primeras columnas del fichero no son relevantes para los cálculos que queremos hacer, por lo que no se leerán. Sin embargo hay que tener en cuenta que Matlab interpreta que un fichero de texto empieza en la fila 0 columna 0. Por tanto, el primer dato a leer es 5.9, que se encuentra en la fila 1 columna 2, y el último dato a leer es 1, que se encuentra en la fila 53 columna 11:

```
M = dlmread('mercurio.txt','\t',[1 2 53 11]);
```

El símbolo `'\t'` indica que los datos están separados por tabulación.

Sólo queremos representar de forma conjunta las variables X_3, X_6, X_7, que son las columnas 1, 4, 5 de la matriz `M`. Así pues construimos una matriz `X` que contenga solamente estas columnas:

```
X = [M(:,1) M(:,4:5)];
det(cov(X,1))
trace(cov(X,1))
plotmatrix(X)
```

La Figura 2.2 muestra la dispersión de las columnas de la matriz `X`.

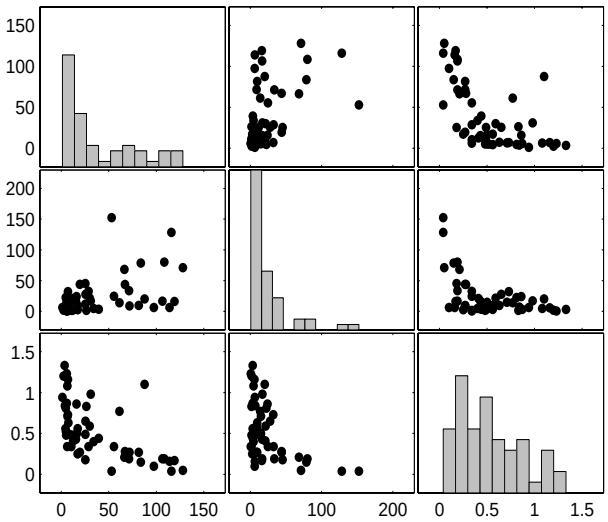


Figura 2.2.
Datos de contaminación por mercurio (Problema 2.3)

Consideremos la siguiente transformación lineal sobre X_3 y X_6:

$$Y_3 = X_3/1000, \quad Y_6 = X_6/1000,$$

que corresponde al cambio de unidades de medida g/l en lugar de mg/l. Y estudiemos ahora la dispersión entre Y_3, Y_6, X_7.

```
Y=[X(:,1)/1000 X(:,2)/1000 X(:,3)];
det(cov(Y,1))
trace(cov(Y,1))
plotmatrix(Y)
```

La Figura 2.3 muestra la dispersión entre las columnas de la matriz `Y`. Observad que si no se tienen en cuenta las unidades de medida, las formas de las nubes de puntos entre las Figuras 2.2 y 2.3 son muy parecidas.
Consideremos ahora las siguientes transformaciones no lineales sobre X_3, X_6 y X_7:

$$W_3 = \log(X_3), \quad W_6 = \log(X_6), \quad W_7 = \sqrt{X_7},$$

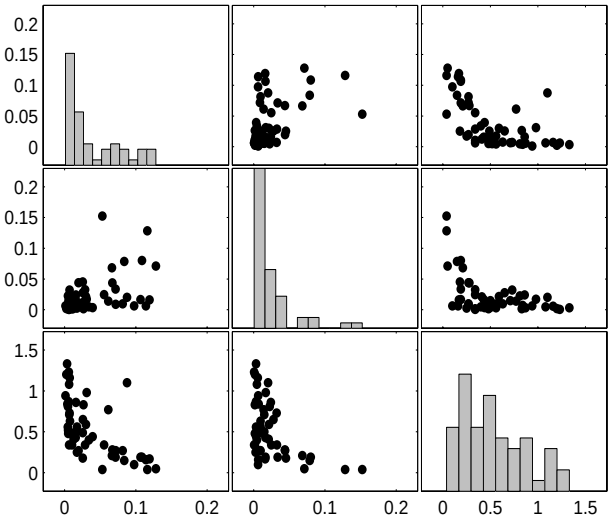


Figura 2.3.
Datos de contaminación por mercurio. Transformaciones lineales (Problema 2.3)

que intentan simetrizar los datos, y estudiemos la dispersión entre ellas:

```
W=[log(X(:,1)) log(X(:,2)) sqrt(X(:,3))];
det(cov(W,1))
trace(cov(W,1))
plotmatrix(W)
```

La Figura 2.4 muestra la dispersión entre las columnas de la matriz `W`.
La Tabla 2.3 resume las medidas de dispersión global para las tres matrices `X`, `Y`, `W`:

Tabla 2.3.
Medidas de dispersión global para las matrices del Problema 2.3

matriz	$tr(\mathbf{S})$	$\det(\mathbf{S})$
`X`	$2.3638e+003$	$6.9503e+004$
`Y`	0.1165	$6.9503e-008$
`W`	3.1223	0.0490

(b) Hemos elegido las transformaciones $\log(X_3)$ y $\sqrt{X_7}$. El código que dibuja los histogramas tridimensionales de la Figura 2.5 se detalla a continuación (Observación: la función `hist3` de Matlab sólo está disponible en la Toolbox Statistics de la versión 7 y superiores). Suponemos que la matriz `M` es la misma que en (a).

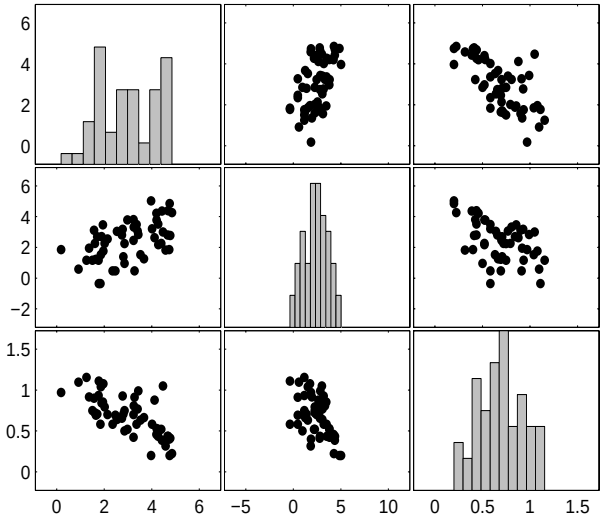


Figura 2.4.
Datos de contaminación por mercurio. Transformaciones no lineales (Problema 2.3)

```
X = M(:,[5,1]);
figure(1)
hist3(X)
ylabel('x_3=alcalinidad')
xlabel('x_7=mercurio')
view(50,50)

Y = [sqrt(X(:,1)),log(X(:,2))] ;
figure(2)
hist3(Y)
ylabel('log(x_3)')
xlabel('x_7^{1/2}')
view(50,50)
```

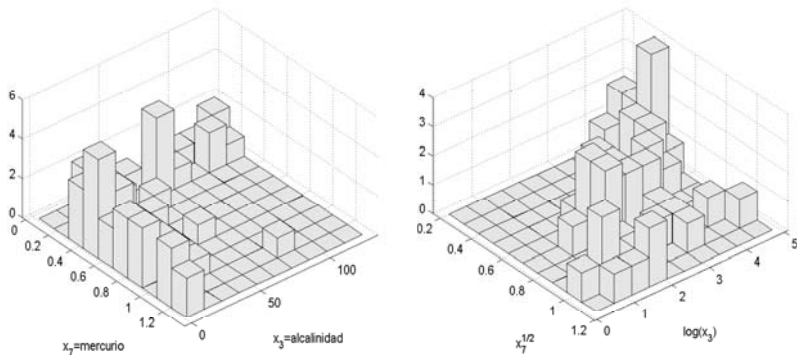


Figura 2.5.
Datos de contaminación por mercurio. Histograma tridimensional (Problema 2.3)

PROBLEMA 2.4

Considérese la muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *de vectores de* $\mathbb{R}^p$*. Pruébese que la matriz de covarianzas*

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

se puede expresar como

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\,\mathbf{x}_i' - \bar{\mathbf{x}}\,\bar{\mathbf{x}}'.$$

SOLUCIÓN

Utilizando la propiedad distributiva de la multiplicación de matrices y que la traspuesta de la suma es la suma de las traspuestas, tenemos que

$$\begin{aligned}
\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' &= \sum_{i=1}^{n}[\mathbf{x}_i(\mathbf{x}_i - \bar{\mathbf{x}})' - \bar{\mathbf{x}}(\mathbf{x}_i - \bar{\mathbf{x}})'] \\
&= \sum_{i=1}^{n}(\mathbf{x}_i\mathbf{x}_i' - \mathbf{x}_i\bar{\mathbf{x}}' - \bar{\mathbf{x}}\mathbf{x}_i' + \bar{\mathbf{x}}\bar{\mathbf{x}}') \\
&= \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' - \sum_{i=1}^{n}\mathbf{x}_i\bar{\mathbf{x}}' - \bar{\mathbf{x}}\sum_{i=1}^{n}\mathbf{x}_i' + \sum_{i=1}^{n}\bar{\mathbf{x}}\bar{\mathbf{x}}' \\
&= \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' - n\bar{\mathbf{x}}\bar{\mathbf{x}}' - n\bar{\mathbf{x}}\bar{\mathbf{x}}' + n\bar{\mathbf{x}}\bar{\mathbf{x}}'.
\end{aligned}$$

PROBLEMA 2.5

Considérese la matriz de datos

$$\begin{pmatrix}
-2 & 1 & 4 \\
3 & 0 & -1 \\
5 & 1 & 2 \\
-1 & 3 & 6 \\
2 & -7 & 4 \\
-1 & 0 & -1
\end{pmatrix},$$

que recoge $n = 6$ *observaciones de un vector aleatorio* $\mathbf{X} = (X_1, X_2, X_3)'$*.*

(a) *Calcúlense el vector de medias* $\bar{\mathbf{x}}$ *y la matriz de covarianzas muestrales* $\mathbf{S}_x$*.*

(b) *Calcúlese la matriz de covarianzas muestrales de los datos estandarizados a media cero y varianza unidad.*

(c) *Sea el vector aleatorio* $\mathbf{Y} = (Y_1, Y_2)$*, donde* $Y_1 = -X_1 + 2X_2 - X_3$ *e* $Y_2 = X_1 + X_2$*. Calcúlense el vector de medias* $\bar{\mathbf{y}}$ *y la matriz de covarianzas muestrales* $\mathbf{S}_y$ *de* $\mathbf{Y}$*. Calcúlese la matriz de observaciones de* $\mathbf{Y}$ *mediante una operación matricial en la que aparezca la matriz de datos de* $\mathbf{X}$*.*

(d) *Calcúlese la matriz de covarianzas del vector aleatorio* $\mathbf{Z} = (Z_1, Z_2)$*, donde* $Z_1 = Y_1/\sqrt{6}$ *y* $Z_2 = Y_2/\sqrt{2}$*.*

(e) *Calcúlense las matrices de correlaciones de* $\mathbf{X}$*,* $\mathbf{Y}$*,* $\mathbf{Z}$ *y de la matriz de datos obtenida en el apartado (b).*

SOLUCIÓN

(a) El vector de medias muestrales de $\mathbf{X}$ es

$$\overline{\mathbf{x}} = \frac{1}{6}\left(\sum_{i=1}^{6} x_{1i}, \sum_{i=1}^{6} x_{2i}, \sum_{i=1}^{6} x_{3i}\right)' = (1, -0.33, 2.33)'.$$

La matriz de covarianzas muestrales de $\mathbf{X}$ es

$$\mathbf{S}_x = \begin{pmatrix} 6.33 & -2.0000 & -2.0000 \\ -2.00 & 9.8889 & 0.1111 \\ -2.00 & 0.1111 & 6.8889 \end{pmatrix}.$$

A continuación indicamos las instrucciones en Matlab que sirven para calcular estos estadísticos. Sea `X` la matriz de datos, que supondremos que ya tenemos introducida, y sean `m` el vector (fila) de medias, `H` la matriz de centrado y `Sx` la matriz de covarianzas . Entonces

```
[n,p] = size(X);
m = ones(n,1)'*X/n;
H = eye(n)-ones(n,n)/n;
Sx = X'*H*X/n;
```

Las instrucciones internas de Matlab `m=mean(X)` y `Sx=cov(X,1)` proporcionan los mismos resultados.

(b) Sean $\mathbf{H}$ la matriz de centrado , $\mathbf{X}_n$ la matriz de datos y $\mathbf{D}_x = \text{diag}(s_{11}, s_{22}, s_{33})$ la matriz diagonal que contiene la diagonal de $\mathbf{S}_x$. Entonces la matriz de datos estandarizados es

$$\mathbf{H}\mathbf{X}_n\mathbf{D}^{-1/2} = \begin{pmatrix} -1.1921 & 0.4240 & 0.6350 \\ 0.7947 & 0.1060 & -1.2700 \\ 1.5894 & 0.4240 & -0.1270 \\ -0.7947 & 1.0600 & 1.3970 \\ 0.3974 & -2.1200 & 0.6350 \\ -0.7947 & 0.1060 & -1.2700 \end{pmatrix},$$

con matriz de covarianzas:

$$\mathbf{S}_{x_0} = \begin{pmatrix} 1.0000 & -0.2527 & -0.3028 \\ -0.2527 & 1.0000 & 0.0135 \\ -0.3028 & 0.0135 & 1.0000 \end{pmatrix}.$$

Sean `H`, `n` y `p` los calculados en (a). Entonces $\mathbf{S}_{x_0}$ se obtiene mediante:

```
d = sqrt(diag(Sx));
Std = ones(n,1)*d';
X0 = (H*X)./Std;
Sx0 = cov(X0,1);
```

(c) Observemos que el vector $\mathbf{Y}$ se expresa como $\mathbf{Y} = \mathbf{X}\,\mathbf{C}'$, siendo

$$\mathbf{C} = \begin{pmatrix} -1 & 2 & -1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Es decir, $\mathbf{Y}$ es una combinación lineal de $\mathbf{X}$. Por tanto,

$$\overline{\mathbf{y}} = \mathbf{C}\,\overline{\mathbf{x}} = \begin{pmatrix} -1 & 2 & -1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ -0.33 \\ 2.33 \end{pmatrix} = \begin{pmatrix} -4 \\ 0.67 \end{pmatrix}$$

y

$$\mathbf{S}_y = \mathbf{C}\,\mathbf{S}_x\,\mathbf{C}' = \begin{pmatrix} 56.33 & 13.33 \\ 13.33 & 12.22 \end{pmatrix}.$$

Instrucciones en MATLAB:

```
C = [-1 2 -1; 1 1 0];
Y = X*C';
my = m*C';
Sy = C*Sx*C';
```

La primera instrucción calcula los valores observados de $\mathbf{Y}$. Podéis comprobar que mediante `mean(Y)` y `cov(Y,1)` se llega al mismo resultado.

(d) Observemos que el vector $\mathbf{Z}$ se escribe como $\mathbf{Z} = \mathbf{X}\,\mathbf{D}'$, donde

$$\mathbf{D} = \begin{pmatrix} -1/\sqrt{6} & 2/\sqrt{6} & -1/\sqrt{6} \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \end{pmatrix},$$

cuyas filas coinciden con las filas de la matriz $\mathbf{C}$ estandarizadas a norma unidad. Procediendo como en el apartado (b), obtenemos

$$\mathbf{S}_z = \mathbf{D}'\,\mathbf{S}_x\,\mathbf{D} = \begin{pmatrix} 9.39 & 3.85 \\ 3.85 & 6.11 \end{pmatrix}.$$

En Matlab escribiremos:

```
D = [-1/sqrt(6) 2/sqrt(6) -1/sqrt(6)
    1/sqrt(2) 1/sqrt(2) 0];
Z = X*D';
Sz = D*Sx*D';
```

(e) Utilizaremos las mismas instrucciones que en el apartado (c) del Problema 2.2. Si llamamos `Rx`, `Ry` y `Rz` a las matrices de correlaciones de $\mathbf{X}$, $\mathbf{Y}$ y $\mathbf{Z}$, y `Rx0` a la matriz de correlaciones de los datos estandarizados, entonces:

```
dx = (diag(Sx)).^(-0.5);
Rx = diag(dx)*Sx*diag(dx);

dx0 = (diag(Sx0)).^(-0.5);
Rx0 = diag(dx0)*Sx0*diag(dx0);

dy = (diag(Sy)).^(-0.5);
Ry = diag(dy)*Sy*diag(dy);

dz = (diag(Sz)).^(-0.5);
Rz = diag(dz)*Sz*diag(dz);
```

Observad que las matrices de correlaciones de $\mathbf{X}$ y de los datos estandarizados coinciden con la matriz de covarianzas de éstos últimos, y que las matrices de correlaciones de $\mathbf{Y}$ y de $\mathbf{Z}$ también coinciden. Comprobad que utilizando la instrucción interna de Matlab `Rx=corrcoef(X)` se llega a los mismos resultados.

PROBLEMA 2.6

Consideremos las $n = 5$ observaciones

$$\begin{pmatrix} 1 & 6 \\ 3 & 8 \\ -2 & 7 \\ 5 & -3 \\ 2 & 0 \end{pmatrix},$$

de un vector aleatorio $\mathbf{X} = (X_1, X_2)'$. Definimos las combinaciones lineales $\mathbf{c}'\mathbf{X}$ y $\mathbf{b}'\mathbf{X}$ donde $\mathbf{c} = (-2, 1)'$ y $\mathbf{b} = (-1, 3)'$.

(a) *Calculando los valores observados de las combinaciones lineales en cada una de las filas de la matriz de datos, obténganse las medias, las varianzas y la covarianza entre $\mathbf{c}'\mathbf{X}$ y $\mathbf{b}'\mathbf{X}$.*

(b) *Obténganse los estadísticos pedidos en (a), pero utilizando las expresiones matriciales que relacionan los momentos muestrales de una combinación lineal con aquéllos ($\bar{\mathbf{x}}$ y $\mathbf{S}$) de $\mathbf{X}$.*

(c) *Obténgase el vector de medias muestral del vector aleatorio $(X_1^2, X_2^2)'$.*

SOLUCIÓN

(a) Los valores observados de las combinaciones $\mathbf{c}'\mathbf{X}$ y $\mathbf{b}'\mathbf{X}$ vienen dados por `Datosc` y `Datosb` respectivamente. La media muestral de los valores observados de $\mathbf{c}'\mathbf{X}$ es `mc` y la de los valores observados de $\mathbf{b}'\mathbf{X}$ es `mb`. La varianza muestral de $\mathbf{c}'\mathbf{X}$ es `vc` y la de $\mathbf{b}'\mathbf{X}$ es `vb`. La covarianza entre $\mathbf{c}'\mathbf{X}$ y $\mathbf{b}'\mathbf{X}$ es `covbc(1,2)`.

```
X = [ 1  6 ; 3  8 ; -2  7 ; 5 -3 ; 2  0] ;
b = [-1 ; 3] ;  c = [-2 ; 1];

Datosb = X*b ;   Datosc = X*c ;

mb = mean(Datosb) ;   mc = mean(Datosc) ;

vb = var(Datosb,1) ;  vc = var(Datosc,1) ;
covbc = cov(Datosb,Datosc,1) ;
```

(b) El vector de medias muestrales de $\mathbf{X}$ es $\bar{\mathbf{x}} = (1.8, 3.6)'$ y su matriz de covarianzas es

$$\mathbf{S} = \begin{pmatrix} 5.36 & -6.28 \\ -6.28 & 18.64 \end{pmatrix}.$$

La media muestral de $\mathbf{c}'\mathbf{X}$ es

$$\mathbf{c}'\bar{\mathbf{x}} = (-2, 1)\begin{pmatrix} 1.8 \\ 3.6 \end{pmatrix} = 0$$

y, análogamente, la media muestral de $\mathbf{b}'\mathbf{X}$ es $\mathbf{b}'\bar{\mathbf{x}} = 9$.
La varianza muestral de $\mathbf{c}'\mathbf{X}$ es $\mathbf{c}'\mathbf{S}\mathbf{c} = 65.2$ y la de $\mathbf{b}'\mathbf{X}$ es $\mathbf{b}'\mathbf{S}\mathbf{b} = 210.8$. La covarianza muestral entre $\mathbf{c}'\mathbf{X}$ y $\mathbf{b}'\mathbf{X}$ es $\mathbf{c}'\mathbf{S}\mathbf{b} = \mathbf{b}'\mathbf{S}\mathbf{c} = 110.6$. A continuación se pueden ver las instrucciones de Matlab que hacen estos cálculos.

```
m = mean(X) ;   S = cov(X,1) ;
mb = b' * m' ;   mc = c' * m' ;
vb = b' * S * b ;   vc = c' * S * c ;
covbc12 = b' * S * c ;
```

(c) El vector de medias muestral de $(X_1^2, X_2^2)'$ viene dado por

$$\begin{pmatrix} \frac{1}{4}\sum_{i=1}^{4} x_{i1}^2 \\ \frac{1}{4}\sum_{i=1}^{4} x_{i2}^2 \end{pmatrix} = \begin{pmatrix} 8.6 \\ 31.6 \end{pmatrix},$$

siendo x_{ij} el elemento (i, j) de la matriz de datos $\mathbf{X}$. Para calcularlo con Matlab, escribimos:

```
Y = X.^2 ;
my = mean(Y) ;
```

Otra posibilidad es recordar que la varianza muestral correspondiente a X_1, la primera componente de $\mathbf{X}$, es:

$$s_{11} = \frac{1}{4}\sum_{i=1}^{4} x_{i1}^2 - \bar{x}_1^2.$$

Por tanto,

$$\frac{1}{4}\sum_{i=1}^{4} x_{i1}^2 = s_{11} + \bar{x}_1^2 = 5.36 + 1.8^2 = 8.6.$$

Análogamente, si s_{22} denota la varianza muestral de X_2, tenemos que

$$\frac{1}{4}\sum_{i=1}^{4} x_{i2}^2 = s_{22} + \bar{x}_2^2 = 18.64 + 3.6^2 = 31.6.$$

PROBLEMA 2.7

Un biólogo recoge medidas (en mm.) de los cráneos en dos especies, A y B, de ratones. Concretamente observa tres variables X_1, X_2 y X_3 en un conjunto de ratones de los cuales $n_A = 50$ son de la especie A y los restantes $n_B = 60$ son de la especie B.

(a) *Denotemos por $\mathbf{X}_A$ la matriz de datos observados en la especie A. Si*

$$\mathbf{X}_A'\mathbf{1}_{50} = (25.5, 14.1, 11.3)'$$

y

$$\mathbf{X}_A'\mathbf{X}_A = \begin{pmatrix} 40.2 & 10.9 & 15.6 \\ 10.9 & 13.7 & 14.5 \\ 15.6 & 14.5 & 20.1 \end{pmatrix},$$

calcúlense el vector de medias $\bar{\mathbf{x}}_A$ y la matriz de covarianzas $\mathbf{S}_A$ correspondientes a esta especie.

(b) *Denotemos por $\mathbf{X}_B$ la matriz de observaciones de la especie B. Si*

$$\mathbf{X}_B'\mathbf{1}_{60} = (26.3, 15.5, 10.0)'$$

y

$$\mathbf{X}_B'\mathbf{X}_B = \begin{pmatrix} 50.7 & 32.6 & 24.8 \\ 32.6 & 29.0 & 12.6 \\ 24.8 & 12.6 & 35.8 \end{pmatrix},$$

calcúlense las medias muestrales $\bar{\mathbf{x}}_B$ y la matriz de covarianzas $\mathbf{S}_B$ de la especie B.

(c) *Calcúlense las medias muestrales $\bar{\mathbf{x}}$ y la matriz de covarianzas $\mathbf{S}$ para la totalidad de los $n = 110$ ratones.*

SOLUCIÓN

(a) Supongamos que los datos están ordenados de manera que los 50 primeros son los de la especie A y los 60 últimos son de la especie B. Entonces tenemos que

$$\mathbf{X}_A'\mathbf{1}_{50} = \begin{pmatrix} \sum_{i=1}^{50} x_{i1} \\ \sum_{i=1}^{50} x_{i2} \\ \sum_{i=1}^{50} x_{i3} \end{pmatrix}.$$

Por tanto,

$$\bar{\mathbf{x}}_A = \frac{1}{50}\mathbf{X}_A'\mathbf{1}_{50} = \begin{pmatrix} 0.5 \\ 0.3 \\ 0.2 \end{pmatrix}.$$

Por otro lado, observemos que

$$\mathbf{X}_A'\mathbf{X}_A = \begin{pmatrix} \sum_{i=1}^{50} x_{i1}^2 & \sum_{i=1}^{50} x_{i1}x_{i2} & \sum_{i=1}^{50} x_{i1}x_{i3} \\ \sum_{i=1}^{50} x_{i1}x_{i2} & \sum_{i=1}^{47} x_{i2}^2 & \sum_{i=1}^{50} x_{i2}x_{i3} \\ \sum_{i=1}^{50} x_{i1}x_{i3} & \sum_{i=1}^{47} x_{i2}x_{i3} & \sum_{i=1}^{50} x_{i3}^2 \end{pmatrix}$$

y

$$\bar{\mathbf{x}}_A\bar{\mathbf{x}}_A' = \begin{pmatrix} 0.3 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix}.$$

Luego

$$\mathbf{S}_A = \frac{1}{50}\mathbf{X}_A'\mathbf{X}_A - \bar{\mathbf{x}}_A\bar{\mathbf{x}}_A' = \begin{pmatrix} 0.5 & 0.1 & 0.2 \\ 0.1 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.4 \end{pmatrix}.$$

Las instrucciones para hacer estos cálculos en Matlab son las siguientes:

```
nA = 50 ;
DatoA1 = [ 25.5,14.1,11.3 ]';
DatoA2 = [ 40.2  10.9  15.6
           10.9  13.7  14.5
           15.6  14.5  20.1 ] ;
MediaA = DatoA1 / nA ;
SA = DatoA2 / nA - MediaA * MediaA' ;
```

(b) Este apartado se resuelve de forma análoga al anterior y los resultados son:

$$\bar{\mathbf{x}}_B = \begin{pmatrix} 0.4 \\ 0.3 \\ 0.2 \end{pmatrix}$$

y

$$\mathbf{S}_B = \begin{pmatrix} 0.7 & 0.4 & 0.3 \\ 0.4 & 0.4 & 0.2 \\ 0.3 & 0.2 & 0.6 \end{pmatrix}.$$

(c) El vector de medias viene dado por

$$\bar{\mathbf{x}} = \frac{1}{110}\begin{pmatrix} \sum_{i=1}^{110} x_{i1} \\ \sum_{i=1}^{110} x_{i2} \\ \sum_{i=1}^{110} x_{i3} \end{pmatrix} = \frac{1}{110}(\mathbf{X}'_A\mathbf{1}_{50} + \mathbf{X}'_B\mathbf{1}_{60}) = \begin{pmatrix} 0.5 \\ 0.3 \\ 0.2 \end{pmatrix}.$$

La matriz de covarianzas es

$$\mathbf{S} = \frac{1}{110}\mathbf{X}'\mathbf{X} - \bar{\mathbf{x}}\,\bar{\mathbf{x}}',$$

donde

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_A \\ \mathbf{X}_B \end{pmatrix},$$

por tanto,

$$\mathbf{S} = \frac{1}{110}(\mathbf{X}'_A\,\mathbf{X}_A + \mathbf{X}'_B\,\mathbf{X}_B) - \bar{\mathbf{x}}\,\bar{\mathbf{x}}' = \begin{pmatrix} 0.6 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.2 \\ 0.3 & 0.2 & 0.5 \end{pmatrix}.$$

Con Matlab

```
n = nA + nB ;
Media = (DatoA1 + DatoB1)/n ;
S = (DatoA2 + DatoB2)/n - Media*Media' ;
```

PROBLEMA 2.8

La Tabla 2.4 contiene 10 observaciones de un vector $\mathbf{X} = (X_1, X_2, X_3, X_4)'$*, donde* X_1 *= Longitud de cabeza del primer hijo de una familia,* X_2 *= Anchura de cabeza de ese mismo hijo,* X_3 *= Longitud de cabeza del segundo hijo de la misma familia y* X_4 *= Anchura de cabeza de este segundo hijo (Fuente: Frets 1921). Divídase* $\mathbf{X}$ *de la siguiente manera:*

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \hline X_3 \\ X_4 \end{pmatrix} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \hline \mathbf{X}^{(2)} \end{pmatrix}.$$

(a) *Para* $\mathbf{X}^{(1)}$ *y* $\mathbf{X}^{(2)}$ *calcúlense, respectivamente, estimaciones de los vectores de esperanzas,* $E(\mathbf{X}^{(1)})$ *y* $E(\mathbf{X}^{(2)})$*, de las matrices de covarianzas,* $Var(\mathbf{X}^{(1)})$ *y* $Var(\mathbf{X}^{(2)})$*, y también de la matriz de covarianzas cruzadas* $Cov(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$.

(b) *Dadas las matrices*

$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \quad y \quad \mathbf{B} = \begin{pmatrix} -3 & 2 \end{pmatrix},$$

calcúlense estimaciones de $E(\mathbf{A}\mathbf{X}^{(1)})$*,* $Var(\mathbf{B}\mathbf{X}^{(2)})$ *y* $Cov(\mathbf{A}\mathbf{X}^{(1)}, \mathbf{B}\mathbf{X}^{(2)})$.

Tabla 2.4.
Dimensiones de cabeza de dos hermanos (Frets 1921)

Primer hijo		Segundo hijo	
Longitud cabeza	Ancho cabeza	Longitud cabeza	Ancho cabeza
191	155	179	145
195	149	201	152
181	148	185	149
183	153	188	149
176	144	171	142
208	157	192	152
189	150	190	149
197	159	189	152
188	152	197	159
192	150	187	151

SOLUCIÓN

(a) Para calcular las estimaciones de los vectores de medias utilizaremos el siguiente código Matlab, en el que suponemos que ya hemos introducido la matriz `X` de datos de dimensión 10×4:

```
MediaHijo1 = mean(X(:,[1,2]))
MediaHijo2 = mean(X(:,[3,4]))
```

o, alternativamente, también podemos hacer:

```
Media = mean(X) ;
MediaHijo1 = Media(1,[1,2])
MediaHijo2 = Media(1,[3,4])
```

Los resultados que se obtienen son $\bar{\mathbf{x}}^{(1)} = (190, 151.7)'$, $\bar{\mathbf{x}}^{(2)} = (187.9, 150)'$. Las estimaciones de las matrices de covarianzas se calculan mediante:

```
S = cov(X,1) ;
VarianzasHijo1 = S([1,2],[1,2])
VarianzasHijo2 = S([3,4],[3,4])
CovHijo1Hijo2 = S([1,2],[3,4])
```

y los resultados son:

$$\mathbf{S}^{(1)} = \begin{pmatrix} 73.4 & 26.6 \\ & 18.0 \end{pmatrix},$$
$$\mathbf{S}^{(2)} = \begin{pmatrix} 65.1 & 29.8 \\ & 18.6 \end{pmatrix},$$
$$\mathbf{S}^{(1,2)} = \begin{pmatrix} 37.8 & 16.8 \\ & 7.0 \end{pmatrix}.$$

(b) Las estimaciones de la esperanza $E(\mathbf{AX}^{(1)})$ y de la varianza $Var(\mathbf{BX}^{(2)})$ son, respectivamente, $\mathbf{A}\,\bar{\mathbf{x}}^{(1)} = (38.3, 341.7)'$ y $\mathbf{BS}^{(2)}\mathbf{B}' = 302.6$. Por último, la estimación de la covarianza cruzada $Cov(\mathbf{AX}^{(1)}, \mathbf{BX}^{(2)})$ es $\mathbf{AS}^{(1,2)}\mathbf{B}' = (-61.7900, -97.8)'$. Una vez introducidas en Matlab las transformaciones lineales `A` y `B`, las instrucciones que calculan estos resultados son:

```
AMediaHijo1 = A * MediaHijo1'
BVarianzasHijo2 = B * VarianzasHijo2 * B'
CovAHijo1BHijo2 = A * CovHijo1Hijo2 * B'
```

PROBLEMA 2.9

Considérese el vector $\mathbf{Y}$ *formado sólo por las dos componentes* X_1 *y* X_2 *del Problema 2.8 centradas respecto de la media muestral* $(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2)'$*. Represéntense las observaciones del vector* $\mathbf{Y}$*,* $\mathbf{y}_i$*, para* $i = 1, \ldots, 10$*, mediante un diagrama de dispersión. A continuación considérese el vector* $\mathbf{a} = (15, 8)'$ *y, sobre el diagrama de dispersión, trácese (a mano o con Matlab) la recta de dirección* $\mathbf{a}$ *que pasa por el origen. Márquese sobre esta recta la proyección ortogonal de* $\mathbf{y}_i$ *sobre* $\mathbf{a}$*, para* $i = 1, \ldots, 10$*, y denótese por* l_i *cada una de estas longitudes. Calcúlese la varianza muestral de las longitudes* l_i*,* $i = 1, \ldots, 10$*. Si, en lugar del vector* $\mathbf{a}$*, se considera el vector* $\mathbf{b} = (15, -15)'$*, ¿qué cambios se observan?*

SOLUCIÓN

Suponemos ya introducida en Matlab la matriz de datos `X` que contiene las dos primeras columnas de la Tabla 2.4. Para realizar el diagrama de dispersión escribimos:

```
Media = mean(X) ;
[n,p] = size(X) ;
Y = X-ones(n,1)*Media ;
plot(Y(:,1),Y(:,2),'ok','MarkerFaceColor','k','MarkerSize',7)
xlabel('y_1')
ylabel('y_2')
axis([-15 20 -15 20])
```

El resultado son los círculos rellenos de la Figura 2.6. La longitud l_i de la proyección de $\mathbf{y}_i = (y_{i1}, y_{i2})'$ sobre $\mathbf{a} = (a_1, a_2)'$ viene dada por $l_i = c_1\, y_{i1} + c_2\, y_{i2}$ (véase el Problema 1.2), siendo $\mathbf{c} = (c_1, c_2)' = \mathbf{a}/\|\mathbf{a}\|$. Análogamente, se obtendrían las longitudes para las proyecciones de $\mathbf{y}_i$ sobre el vector $\mathbf{b}$. Las siguientes instrucciones permiten realizar los cálculos en Matlab:

```
a = [15;8]; b = [15;-15];
c = a/norm(a);   d = b/norm(b);
La = Y*c;   Lb = Y*d;
var(La)
var(Lb)
```

La varianza resultante de las proyecciones sobre el vector $\mathbf{a}$ es `var(La)=92.47`, mientras que proyectando sobre $\mathbf{b}$ la varianza es `var(Lb)=21.23`, que es bastante menor. Para añadir estas proyecciones al gráfico anterior, escribimos:

```
Ya = La*c';   Yb = Lb*d';
hold on
plot(Ya(:,1),Ya(:,2),'^b')
plot(Yb(:,1),Yb(:,2),'*r')
```

Las proyecciones sobre $\mathbf{a}$ y $\mathbf{b}$ aparecen representadas en la Figura 2.6 con triángulos y estrellas, respectivamente. En las técnicas de análisis multivariante que se exponen a partir del Capítulo 4 es importante tener en cuenta las consecuencias de elegir distintas direcciones sobre las que proyectar los datos.

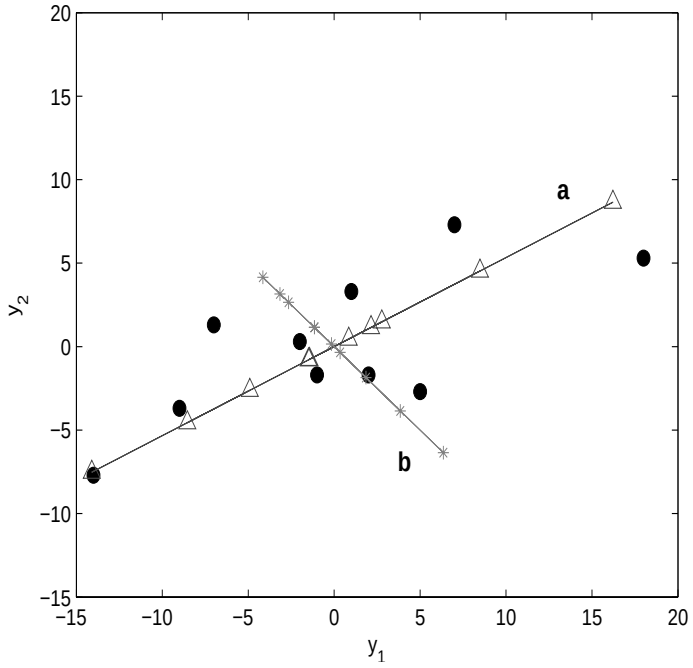


Figura 2.6.
Proyección ortogonal de observaciones (Problema 2.9)

CAPÍTULO 3

Distribuciones multivariantes

En este capítulo se presentan diversos conceptos y herramientas estadísticas útiles para describir la distribución de un vector aleatorio: vector de medias, matriz de covarianzas, función de densidad, ... A lo largo del tema se hace hincapié en las distintas propiedades de los momentos de un vector aleatorio (por ejemplo, bajo transformaciones lineales del mismo).
También se trabaja con la distribución más importante en el contexto multivariante, la distribución normal. Con diversos ejercicios se repasan las propiedades que caracterizan esta distribución, entre otras que es el límite al que converge la media muestral (Teorema Central del Límite). Por último, se consideran algunas otras distribuciones, como la T^2 de Hotelling , la ley de Wishart o la Lambda de Wilks, que resultan esenciales a la hora de hacer inferencia sobre datos multivariados.

PROBLEMA 3.1

Sea $\mathbf{X}$ *un vector aleatorio* p*-dimensional de media* $\boldsymbol{\mu}$ *y matriz de varianzas-covarianzas* $\mathbf{I}$ *(la matriz identidad de dimensión* $p \times p$*). Dada una matriz cuadrada de orden* p*,* $\mathbf{A}$*, considérese la nueva variable* $Y = \mathbf{X}' \mathbf{A} \mathbf{X}$ *y demuéstrese que*

$$E(Y) = tr(\mathbf{A}) + \boldsymbol{\mu}' \mathbf{A} \boldsymbol{\mu} \, .$$

SOLUCIÓN

Si denotamos por $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ y $\mathbf{A} = (a_{ij})_{1 \leq i,j \leq p}$, entonces

$$Y = \mathbf{X}' \mathbf{A} \mathbf{X} = \sum_{i,j=1}^{p} a_{ij} \, X_i \, X_j .$$

Por tanto,

$$E(Y) = \sum_{i,j=1}^{p} a_{ij}\ E(X_i X_j) = \sum_{i=1}^{p} a_{ii}\ E(X_i^2) + \sum_{\substack{i,j=1 \\ i \neq j}}^{p} a_{ij}\ E(X_i X_j).$$

Puesto que la matriz de covarianzas de $\mathbf{X}$ es la identidad, tenemos que $E(X_i^2) = 1 + \mu_i^2$, para $i = 1, \ldots, p$, y también que $E(X_i X_j) = E(X_i)\ E(X_j) = \mu_i\, \mu_j$, para todo $i \neq j$. Entonces

$$\begin{aligned} E(Y) &= \sum_{i=1}^{p} a_{ii}(1 + \mu_i^2) + \sum_{\substack{i,j=1 \\ i \neq j}}^{p} a_{ij}\mu_i\mu_j \\ &= \sum_{i=1}^{p} a_{ii} + \sum_{i,j=1}^{p} a_{ij}\mu_i\mu_j = \text{tr}(\mathbf{A}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}. \end{aligned}$$

PROBLEMA 3.2

Supongamos que X_1, X_2, X_3 son v.a. independientes con varianza unidad. Sean $Y_1 = X_1 + X_2 + X_3$, $Y_2 = X_1 - X_2$ e $Y_3 = X_1 - X_3$. Calcúlense las matrices de varianzas-covarianzas y de correlaciones de $\mathbf{Y} = (Y_1, Y_2, Y_3)'$.

SOLUCIÓN

La matriz de covarianzas de $\mathbf{X}$ es $Var(\mathbf{X}) = \mathbf{I}$, la matriz identidad de dimensión 3×3. Puesto que $\mathbf{Y} = \mathbf{A}\,\mathbf{X}$, siendo

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix},$$

la matriz de varianzas-covarianzas de $\mathbf{Y}$ es

$$Var(\mathbf{Y}) = \mathbf{A}\,\mathbf{A}' = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

y la matriz de correlaciones es

$$Corr(\mathbf{Y}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{pmatrix}.$$

PROBLEMA 3.3

Sea $\mathbf{X} = (X_1, X_2, X_3)'$ *un vector aleatorio tridimensional. Se sabe que el vector* $\mathbf{Y} = (Y_1, Y_2, Y_3)'$ *ha sido generado del siguiente modo:* $\mathbf{Y} = \mathbf{B}\,\mathbf{X}$*, donde*

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 1 \\ -1 & 1 & 1 \\ 0 & 1 & -1 \end{pmatrix}$$

es una matriz no singular. Se sabe también que $E(\mathbf{Y}) = (2, 1, 0)'$ *y que la matriz de covarianzas de* $\mathbf{Y}$ *es*

$$Var(\mathbf{Y}) = \begin{pmatrix} 5 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

(a) *Hállese la covarianza entre* $Z_1 = Y_2 - Y_1$ *y* $Z_2 = Y_1 + Y_3$.

(b) *Calcúlense* $\boldsymbol{\mu} = E(\mathbf{X})$ *y* $\boldsymbol{\Sigma} = Var(\mathbf{X})$*, la matriz de covarianzas de* $\mathbf{X}$.

(c) *Si se define* $\mathbf{Y} = \mathbf{B}\,(\mathbf{X} - \boldsymbol{\mu})$ *¿cuál sería* $E(\mathbf{Y})$*? ¿Cómo es la fórmula para hallar ahora* $Var(\mathbf{Y})$ *a partir de* $\boldsymbol{\mu}$ *y* $\boldsymbol{\Sigma}$*? ¿Depende de* $\boldsymbol{\mu}$*?*

SOLUCIÓN

(a)

$$Cov(Z_1, Z_2) = (-1, 1, 0)\ Var(\mathbf{Y})\,(1, 0, 1)' = -7\,.$$

(b) Sabemos que $\mathbf{Y} = \mathbf{B}\mathbf{X}$, por lo que tendremos que $\mathbf{X} = \mathbf{B}^{-1}\mathbf{Y}$. Por tanto,

$$\boldsymbol{\mu} = \mathbf{B}^{-1}\,E(\mathbf{Y}) = \frac{1}{3}\begin{pmatrix} 2 & -1 & 1 \\ 1 & 1 & 2 \\ 1 & 1 & -1 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

$$\boldsymbol{\Sigma} = \mathbf{B}^{-1}\,Var(\mathbf{Y})(\mathbf{B}^{-1})' = \frac{1}{3}\begin{pmatrix} 10 & 4 & 1 \\ 4 & 3 & 0 \\ 1 & 0 & 3 \end{pmatrix}.$$

(c) Si ahora tenemos $\mathbf{Y} = \mathbf{B}\,(\mathbf{X} - \boldsymbol{\mu})$, esto implica que

$$E(\mathbf{Y}) = \mathbf{B}(\boldsymbol{\mu} - \boldsymbol{\mu}) = \mathbf{0} \quad \text{y} \quad Var(\mathbf{Y}) = \mathbf{B}\ Var(\mathbf{X})\,\mathbf{B}'\,,$$

es decir, la varianza no se ve afectada por traslaciones.

PROBLEMA 3.4

Sea $\mathbf{X}$ *un vector con distribución uniforme en el rectángulo* $[0,2]\times[3,4]$.

(a) *Especifíquese la función de densidad de* $\mathbf{X}$. *Calcúlense* $E(\mathbf{X})$ *y* $Var(\mathbf{X})$.

(b) *Sea* $\mathbf{X}_1,\ldots,\mathbf{X}_{30}$ *una muestra aleatoria simple de* $\mathbf{X}$ *y* $\bar{\mathbf{X}}=\sum_{i=1}^{30}\mathbf{X}_i/30$ *la media muestral correspondiente. Calcúlense* $E(\bar{\mathbf{X}})$ *y* $Var(\bar{\mathbf{X}})$.

(c) *Genérese con Matlab una realización de la muestra del apartado anterior. Calcúlense la media* $\bar{\mathbf{x}}$ *y la matriz de covarianzas muestrales* $\mathbf{S}$. *Dibújese en un gráfico de dispersión la muestra y márquense los puntos* $E(\bar{\mathbf{X}})$ *y* $\bar{\mathbf{x}}$.

(d) *Genérense con Matlab 40 muestras de tamaño 5, calcúlense sus correspondientes medias muestrales y dibújense éstas en un gráfico en el que se marque también* $E(\bar{\mathbf{X}})$. *Repítase este proceso en gráficos distintos para 40 muestras de tamaño 20 y otras 40 de tamaño 50. ¿Qué se observa?*

SOLUCIÓN

(a) La densidad es

$$f(x_1,x_2)=\begin{cases}1/2, & \text{si } \mathbf{x}\in[0,2]\times[3,4],\\ 0, & \text{en otro caso.}\end{cases}$$

El vector de esperanzas de $\mathbf{X}$ es $E(\mathbf{X})=(E(X_1),E(X_2))'$, donde

$$E(X_i)=\int_{-\infty}^{\infty}x_i\,f_i(x_i)\,dx_i\,,$$

para $i=1,2$, siendo f_i la función de densidad marginal de la variable aleatoria X_i. Puesto que X_1 y X_2 son v.a. independientes entre sí y con ley uniforme en los intervalos $[0,2]$ y $[3,4]$, respectivamente, $E(\mathbf{X})=(1,3.5)'$, que es el punto central del rectángulo y $Cov(X_1,X_2)=0$. Por otro lado, $Var(X_i)=E(X_i^2)-E(X_i)^2$, luego

$$Var(\mathbf{X})\simeq\begin{pmatrix}0.33 & 0\\ 0 & 8.83\end{pmatrix}.$$

(b) $E(\bar{\mathbf{X}})=E(\mathbf{X})$ y $Var(\bar{\mathbf{X}})=Var(\mathbf{X})/30$.

(c) El siguiente código resuelve este apartado y genera la Figura 3.1.

```
n = 30 ; p = 2 ;
X = rand(n,p) ; % Muestra de una Unif[0,1]*[0,1]
X = [2*X(:,1),3+X(:,2)] ; % Muestra de Unif[0,2]*[3,4]
m = mean(X) ; % Media muestral
S = cov(X,1) ; % Matriz de varianzas-covarianzas muestrales
```

```
plot(X(:,1),X(:,2),'o','MarkerFaceColor','k',...
     'MarkerEdgeColor','k')
axis([0 2 3 4])
hold on
plot(m(1),m(2),'k*','MarkerSize',8)
hold on
plot(1,3.5,'ko','MarkerSize',8)
```

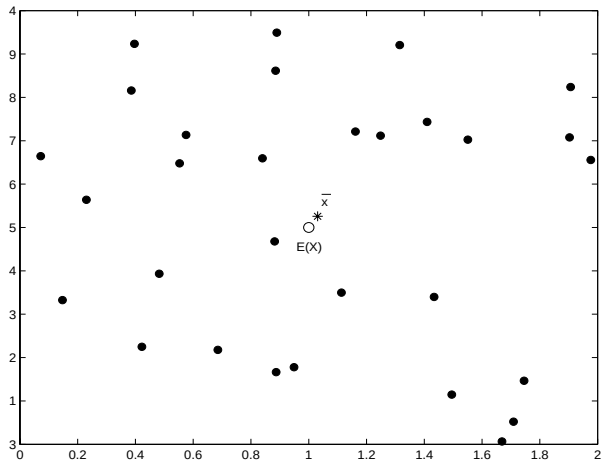
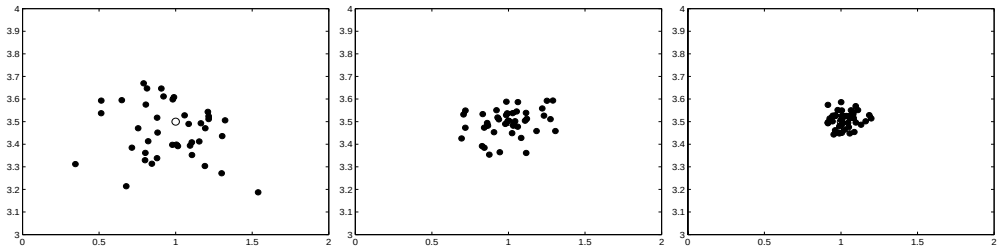


Figura 3.1.
Diagrama de dispersión de muestra uniforme en $[0, 2] \times [3, 4]$ (Problema 3.4)

(d) Se observa que, a mayor tamaño muestral, menor dispersión de la media muestral y mejor estima ésta la esperanza de $\mathbf{X}$. Una propuesta de código es la que sigue, pero retamos al lector a sustituir los bucles `for` por operaciones matriciales (esto reduce el tiempo de ejecución). Los gráficos resultantes se pueden ver en la Figura 3.2.

```
N = 40 ; % Numero de muestras
Vector_n = [5,20,50];
for i = 1:length(Vector_n)
   n = Vector_n(i); % Tamanio muestral
   MatrizMedias = zeros(N,2) ;
   for num =1:N
      X = [2*rand(n,1),3+rand(n,1)] ;
      MatrizMedias(num,:) = mean(X) ;
   end
   figure(i+1)
   plot(MatrizMedias(:,1),MatrizMedias(:,2),'o',...
        'MarkerFaceColor','k','MarkerEdgeColor','k')
   axis([0 2 3 4])
   title(['40 medias muestrales con tamanio muestral ',...
         num2str(n)])
   hold on
   plot(1,3.5,'ko','MarkerSize',8)
   hold off
end
```



(a) (b) (c)

Figura 3.2.
Medias muestrales con tamaño muestral (a) 5, (b) 20, (c) 50 (Problema 3.4)

PROBLEMA 3.5

Sea $\mathbf{X}$ *un vector aleatorio de distribución normal con media* $\boldsymbol{\mu} = (-1, 1, 0)'$ *y matriz de covarianzas*

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

(a) *Hállese la distribución de* $X_1 + 2X_2 - 3X_3$.

(b) *Hállese un vector* $\mathbf{a}_{(2\times 1)}$, *tal que las variables* X_1 *y* $X_1 - \mathbf{a}' \begin{pmatrix} X_2 \\ X_3 \end{pmatrix}$ *sean independientes.*

(c) *Calcúlese la distribución de* X_3 *condicionada a* $X_1 = x_1$ *y* $X_2 = x_2$.

SOLUCIÓN

(a) Se verifica que $Y = X_1 + 2X_2 - 3X_3 = \mathbf{b}'\mathbf{X}$ con $\mathbf{b} = (1, 2, -3)'$. Por tanto,

$$Y \sim N(\mathbf{b}'\boldsymbol{\mu}, \mathbf{b}'\boldsymbol{\Sigma}\mathbf{b}) = N(1, 13).$$

(b) Por la hipótesis de normalidad X_1 y $X_1 - \mathbf{a}' \begin{pmatrix} X_2 \\ X_3 \end{pmatrix}$ son independientes si y sólo si

$$Cov\left(X_1, X_1 - \mathbf{a}' \begin{pmatrix} X_2 \\ X_3 \end{pmatrix}\right) = 0.$$

Por tanto, debemos hallar $\mathbf{a} = (a_1, a_2)'$ tal que se verifique esta última condición. Puesto que

$$Cov\left(X_1, X_1 - \mathbf{a}' \begin{pmatrix} X_2 \\ X_3 \end{pmatrix}\right) = (1, 0, 0)\, \boldsymbol{\Sigma}\, (1, -a_1, -a_2)' = 1 - a_2,$$

deducimos que $a_2 = 1$. Por ejemplo, podemos tomar $\mathbf{a} = (0, 1)'$.

(c) La variable $X_3|X_1 = x_1, X_2 = x_2$ sigue una distribución $N(\mu_c, \Sigma_c)$, donde

$$\begin{aligned}\mu_c &= \mu_3 + \mathit{Cov}\left(X_3, \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right) (\mathit{Var}(X_1, X_2))^{-1} \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\right) \\ &= \mu_3 + (\mathit{Cov}(X_3, X_1), \mathit{Cov}(X_3, X_2))\ (\mathit{Var}(X_1, X_2))^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \\ &= 0 + (1,1)\,\frac{1}{3} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 + 1 \\ x_2 - 1 \end{pmatrix} = \frac{1}{3}\,(3x_1 + x_2 + 2)\,,\end{aligned}$$

$$\begin{aligned}\Sigma_c &= \mathit{Var}(X_3) - \mathit{Cov}\left(X_3, \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right) (\mathit{Var}(X_1, X_2))^{-1}\, \mathit{Cov}\left(X_3, \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right)' \\ &= 2 - (1,1)\,\frac{1}{3} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} (1,1)' = \frac{2}{3}\,.\end{aligned}$$

PROBLEMA 3.6

Sean X_1, X_2 y X_3 tres variables aleatorias con distribución conjunta normal con vector de medias $\boldsymbol{\mu} = (0,0,0)'$ y matriz de varianzas-covarianzas

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

Calcúlese la distribución conjunta de

(a) $Y_1 = X_1 + X_3$ *e* $Y_2 = X_2 + X_3$,

(b) $Z_1 = 3X_1 - 2X_2$, $Z_2 = 2X_1 - X_2 + X_3$ *y* $Z_3 = 3X_3$.

SOLUCIÓN

(a) Sean $\mathbf{X} = (X_1, X_2, X_3)'$ y $\mathbf{A}$ la transformación lineal siguiente:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Entonces $\mathbf{Y} = (Y_1, Y_2)' = \mathbf{A}\,\mathbf{X}$, es una combinación lineal de $\mathbf{X}$ y, por tanto, tiene una distribución normal de parámetros $E(\mathbf{Y}) = \mathbf{A}\,\boldsymbol{\mu} = (0,0)'$ y

$$\mathit{Var}(\mathbf{Y}) = \mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{A}' = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}.$$

(b) Consideremos ahora la transformación lineal dada por la matriz

$$\mathbf{B} = \begin{pmatrix} 3 & -2 & 0 \\ 2 & -1 & 1 \\ 0 & 0 & 3 \end{pmatrix}$$

y observemos que $\mathbf{Z} = (Z_1, Z_2, Z_3)' = \mathbf{B}\,\mathbf{X}$. Por tanto, $\mathbf{Z}$ sigue una distribución normal de media $E(\mathbf{Z}) = \mathbf{0}$ y

$$Var(\mathbf{Z}) = \mathbf{B}\,\mathbf{\Sigma}\,\mathbf{B} = \begin{pmatrix} 17 & 12 & 6 \\ 12 & 10 & 9 \\ 6 & 9 & 18 \end{pmatrix}.$$

PROBLEMA 3.7

Sea $\mathbf{X} = (X_1, X_2, X_3)'$ un vector aleatorio tridimensional que sigue una distribución normal con media $\boldsymbol{\mu} = (1, 0, -2)'$ y matriz de varianzas-covarianzas

$$\mathbf{\Sigma} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 4 & 1 \\ 0 & 1 & 6 \end{pmatrix}.$$

(a) *Escríbase la forma cuadrática $Q(x_1, x_2, x_3)$ del exponente de la densidad del vector aleatorio $\mathbf{X}$.*

(b) *Escríbase la matriz de covarianzas cruzadas entre $\begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$ y X_2.*

(c) *Encuéntrese la correlación entre X_1 y X_3 condicionadas por $X_2 = x_2$.*

(d) *Hállese var$(X_1|X_2 = x_2)$ y compárese con var(X_1).*

SOLUCIÓN

(a) Sea $\mathbf{x} = (x_1, x_2, x_3)'$ un vector de $\mathbb{R}^3$. Puesto que

$$\mathbf{\Sigma}^{-1} = \frac{1}{40}\begin{pmatrix} 23 & 6 & -1 \\ 6 & 12 & -2 \\ -1 & -2 & 7 \end{pmatrix},$$

entonces

$$\begin{aligned} Q(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \\ &= -\frac{1}{80}(x_1 - 1, x_2, x_3 + 2)\begin{pmatrix} 23 & 6 & -1 \\ 6 & 12 & -2 \\ -1 & -2 & 7 \end{pmatrix}\begin{pmatrix} x_1 - 1 \\ x_2 \\ x_3 + 2 \end{pmatrix}. \end{aligned}$$

(b)

$$Cov\left(\begin{pmatrix} X_1 \\ X_3 \end{pmatrix}, X_2\right) = \begin{pmatrix} Cov(X_1, X_2) \\ Cov(X_3, X_2) \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

(c) La distribución de $(X_1, X_3)'$ condicionada por $X_2 = x_2$ es una normal bivariante con matriz de covarianzas

$$\boldsymbol{\Sigma}_c = \begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix} - \begin{pmatrix} -1 \\ 1 \end{pmatrix} \frac{1}{4} (-1, 1) = \frac{1}{4} \begin{pmatrix} 7 & 1 \\ 1 & 23 \end{pmatrix}.$$

Por tanto, la correlación entre X_1 y X_3 condicionadas por $X_2 = x_2$ es

$$corr(X_1, X_3 | X_2 = x_2) = \frac{1/4}{\sqrt{7/4 \cdot 23/4}} = \frac{1}{\sqrt{7 \cdot 23}} \simeq 0.079.$$

(d) A partir de los cálculos realizados en el apartado (c), vemos que

$$var(X_1 | X_2 = x_2) = 7/4\,,$$

que es menor que $var(X_1) = 2$. Esto es razonable puesto que, al condicionar a $X_2 = x_2$, tenemos mayor información acerca de X_1 y su variabilidad disminuye respecto a la distribución sin condicionar.

PROBLEMA 3.8

Sean $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ y $\mathbf{X}_4$ vectores aleatorios independientes con distribución $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde $\boldsymbol{\mu} = (1, 2)'$ y

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 2 \end{pmatrix}.$$

(a) *Hállese la distribución del vector aleatorio*

$$\mathbf{Y} = \frac{1}{4}\mathbf{X}_1 - \frac{1}{4}\mathbf{X}_2 + \frac{1}{4}\mathbf{X}_3 - \frac{1}{4}\mathbf{X}_4.$$

(b) *Escríbase y dibújese (con Matlab) la densidad del vector $\mathbf{Y}$ dado en (a).*

(c) *Calcúlese la correlación ρ correspondiente a la matriz de covarianzas $\boldsymbol{\Sigma}$. Cámbiese el valor de ρ y vuélvase a dibujar la densidad de $\mathbf{Y}$. ¿Qué cambios se observan?*

SOLUCIÓN

(a) El vector $\mathbf{Y}$ sigue una distribución normal bivariante de media

$$E(\mathbf{Y}) = \left(\frac{1}{4} - \frac{1}{4} + \frac{1}{4} - \frac{1}{4}\right)\boldsymbol{\mu} = \mathbf{0}$$

y matriz de covarianzas

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = \left(\left(\frac{1}{4}\right)^2 + \left(-\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(-\frac{1}{4}\right)^2\right)\boldsymbol{\Sigma} = \frac{1}{4}\boldsymbol{\Sigma}.$$

(b) Como $E(\mathbf{Y}) = \mathbf{0}$ la función de densidad de $\mathbf{Y} = (Y_1, Y_2)'$ tiene la expresión

$$f(\mathbf{y}) = \frac{1}{2\pi|\boldsymbol{\Sigma}_{\mathbf{Y}}|^{1/2}} \exp\left(-\frac{1}{2}(y_1, y_2)\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right).$$

Para dibujar la función f (véase la Figura 3.3.a) escribimos:

```
mu = [ 1 ; 2 ] ;
Sigma = [ 1  0.1 ; 0.1  2 ] ;
c = [ 1/4 , -1/4 , 1/4 , -1/4 ] ;
mY = sum(c) * mu ;
SY = sum(c.^2) *Sigma ;

y1 = [-2:0.1:2] ;  y2 = [-2:0.1:2] ;
[Y1,Y2] = meshgrid(y1,y2) ;
[m,n] = size(Y1) ;  f = zeros(m,n) ;
for i = 1:m
  for j=1:n
    y = [ Y1(i,j) ; Y2(i,j) ] ;
    f(i,j) = exp(-0.5*(y-mY)'*inv(SY)*(y-mY))/...
                (2*pi*sqrt(det(SY)));
  end
end
mesh(Y1,Y2,f)
view(-57,40)
xlabel('y_1')
ylabel('y_2')
```

(c) La correlación que nos piden es $\rho = 0.1/\sqrt{2} \simeq 0.071$. Si cambiamos su valor a, por ejemplo, $\rho = 0.8$ sin alterar las varianzas de $\boldsymbol{\Sigma}$, la matriz pasa a ser

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \sqrt{2}\rho \\ \sqrt{2}\rho & 2 \end{pmatrix}.$$

Esta matriz la introduciremos mediante el código

```
rho = 0.8;
NewSigma = zeros(size(Sigma)) ;  NewSigma(2,2) = Sigma(2,2) ;
NewSigma(1,1) = Sigma(1,1) ;
NewSigma(1,2) = sqrt(Sigma(1,1)*Sigma(2,2))*rho ;
NewSigma(2,1) = NewSigma(1,2) ;
```

y el resto se hace de manera análoga al apartado (b). En la Figura 3.3.b se encuentra la representación gráfica de esta nueva densidad del vector $\mathbf{Y}$. Observad cómo varía su forma en función de ρ.

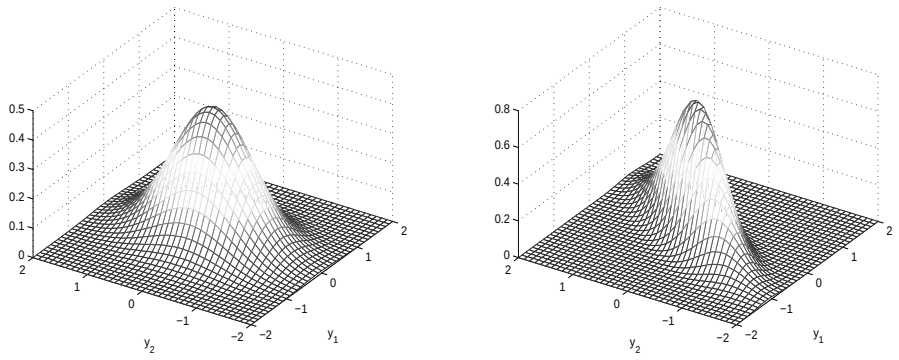


Figura 3.3.
Densidad de un vector normal para (a) $\rho = 0.071$ y (b) $\rho = 0.8$. (Problema 3.8)

PROBLEMA 3.9

Consideremos la muestra

$$\mathbf{X} = \begin{pmatrix} 2 & 6 & -3 \\ -4 & 8 & 7 \\ -2 & 9 & 7 \\ -7 & 8 & 2 \end{pmatrix}$$

de una población $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *con* $\boldsymbol{\mu}$ *y* $\boldsymbol{\Sigma}$ *desconocidos.*

(a) *Calcúlese el estimador de máxima verosimilitud de* $\boldsymbol{\mu}$.

(b) *Calcúlese un estimador insesgado de* $\boldsymbol{\Sigma}$.

(c) *Calcúlese la matriz de varianzas-covarianzas muestrales.*

SOLUCIÓN

(a) El estimador de máxima verosimilitud de $\boldsymbol{\mu}$ es la media muestral $\bar{\mathbf{x}} = (-2.75, 7.75, 3.25)'$.

(b) Si denotamos por $\mathbf{H}$ la matriz de centrado, un estimador insesgado de $\boldsymbol{\Sigma}$ es

$$\tilde{\mathbf{S}} = \frac{1}{n-1}\mathbf{X}'\mathbf{H}\mathbf{X} = \begin{pmatrix} 14.2 & -2.9 & -8.8 \\ -2.9 & 1.6 & 5.4 \\ -8.8 & 5.4 & 22.9 \end{pmatrix}.$$

(c) La matriz de varianzas-covarianzas muestrales es

$$\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{H}\mathbf{X} = \begin{pmatrix} 10.7 & -2.2 & -6.6 \\ -2.2 & 1.2 & 4.1 \\ -6.6 & 4.1 & 17.2 \end{pmatrix}.$$

PROBLEMA 3.10

Sea $\mathbf{X}_1, \ldots, \mathbf{X}_{80}$ una muestra de una población con media $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$.

(a) *¿Cuál es la distribución aproximada de*

$$\bar{\mathbf{X}} = \sum_{i=1}^{80} X_i/80 \ ?$$

(b) *Tómense $N = 200$ muestras de tamaño $n = 80$ de un vector $\mathbf{X} = (X_1, X_2)'$ con distribución uniforme en el cuadrado $[0,1] \times [0,1]$. Calcúlense las medias $\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_N$ de estas muestras y dibújese el histograma correspondiente a las medias, comprobando si se asemeja a una densidad normal.*

SOLUCIÓN

(a) Por el Teorema Central del Límite el vector $\bar{\mathbf{X}}$ sigue aproximadamente una distribución normal de media $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}/75$.

(b) El siguiente código dibuja el histograma pedido.

```
N = 200 ;
n = 80 ;
MatrizMedias = zeros(N,2) ;
for i = 1:N
   muestra = rand(n,2) ;
   MatrizMedias(i,:) = mean(muestra) ;
end
hist3(MatrizMedias) ;
```

Probablemente el histograma obtenido no se parezca excesivamente a una densidad normal salvo en la aparente unimodalidad y simetría (véase la Figura 3.4). Por ello es interesante tomar valores de N y n bastante mayores para comprobar la convergencia a la normal.

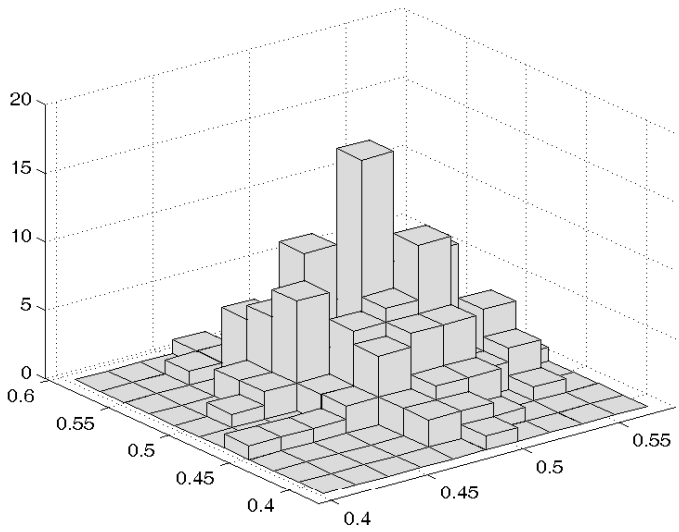


Figura 3.4.
Histograma de medias de una uniforme (Problema 3.10)

PROBLEMA 3.11

Sean X_1, X_2 y X_3 los niveles de solvencia de tres bancos españoles. Supongamos que la distribución conjunta de los tres niveles es $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $\boldsymbol{\mu} = (0.7, 0.8, 0.9)'$ y

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Consideremos un nivel de solvencia medio para los tres bancos que se obtiene mediante el promedio $W = (X_1 + X_2 + X_3)/3$.

(a) *Calcúlese la distribución del nivel de solvencia medio W.*

(b) *Encuéntrese la distribución de $(X_1, X_2)'$ condicionada a que W vale 1.*

(c) *¿Son X_2 y W independientes?*

SOLUCIÓN

(a) Dado que $W = \frac{1}{3}(1,1,1)(X_1, X_2, X_3)'$, se tiene que W sigue una normal de media $\frac{1}{3}(1,1,1)\boldsymbol{\mu} = 0.8$ y varianza $\frac{1}{3^2}(1,1,1)\boldsymbol{\Sigma}(1,1,1)' = \frac{1}{3}$.

(b) Observemos que

$$(X_1, X_2, W)' = \mathbf{A}(X_1, X_2, X_3)',$$

donde

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$

es una combinación lineal de $(X_1, X_2, X_3)'$. Por tanto, $(X_1, X_2, W)'$ sigue una distribución

$$N_3(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'),$$

con

$$\mathbf{A}\boldsymbol{\mu} = (0.7, 0.8, 0.8)'$$

y

$$\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \begin{pmatrix} 2 & -1 & \frac{1}{3} \\ -1 & 2 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$

La distribución de $(X_1, X_2)'|W = 1$ es $N_2(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, con $\boldsymbol{\mu}_c = (0.9, 1)'$ y

$$\boldsymbol{\Sigma}_c = \frac{1}{3}\begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix}.$$

(c) A partir de la expresión obtenida en el apartado (b) para la matriz de covarianzas del vector $(X_1, X_2, W)'$, se tiene que

$$Cov(X_2, W) = 1/3 \neq 0\,,$$

es decir, X_2 y W no son independientes.

PROBLEMA 3.12

Razona si, en tu opinión, los datos que aparecen representados en el diagrama de dispersión múltiple de la Figura 3.5 pueden provenir de una distribución normal multivariante.

SOLUCIÓN

Una propiedad de la normal multivariante es que sus marginales univariantes son también normales. Si la muestra representada en el gráfico proviniera de una población normal, los histogramas de las marginales univariantes deberían exhibir propiedades (como la simetría) propias de una normal y esto no sucede para ninguna de las cinco variables representadas.

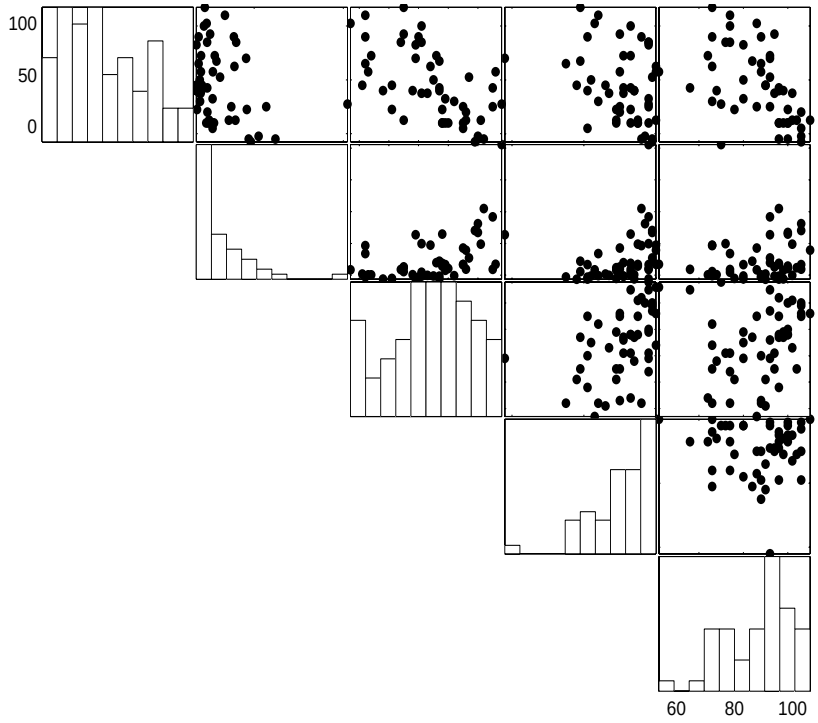


Figura 3.5.
Diagrama de dispersión múltiple (Problema 3.12)

PROBLEMA 3.13

Con algunos programas de ordenador sólo se pueden generar muestras normales univariantes. Supongamos, sin embargo, que deseamos generar una muestra de un vector bidimensional $\mathbf{Y} = (Y_1, Y_2)'$ *con distribución* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, donde*

$$\boldsymbol{\mu} = (\mu_1, \mu_2)',$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\rho \\ \sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\rho & \sigma_{22} \end{pmatrix}$$

y ρ *denota la correlación entre* Y_1 *e* Y_2*. Entonces podemos recurrir al procedimiento que explicamos a continuación.*

(a) *Con la orden* `randn` *de Matlab, que sólo genera observaciones normales univariantes e independientes entre sí, y para un tamaño muestral* n *a elegir, genérese una muestra*

$$\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \tag{3.1}$$

de un vector $\mathbf{X} = (X_1, X_2)'$ *con distribución* $N_2(\mathbf{0}, \mathbf{I})$.

(b) *Ahora consideremos las siguientes transformaciones lineales de* $\mathbf{X}$

$$\begin{aligned} Y_1 &= \mu_1 + \sqrt{\sigma_{11}} X_1 \\ Y_2 &= \mu_2 + \sqrt{\sigma_{22}}(\rho X_1 + \sqrt{1-\rho^2} X_2). \end{aligned} \tag{3.2}$$

Demuéstrese que $\mathbf{Y} = (Y_1, Y_2)'$ *sigue una distribución* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

(c) *Elíjanse unos valores concretos para* $\boldsymbol{\mu}$, σ_{11}, σ_{22} *y* ρ. *Utilizando la combinación lineal (3.2), genérese con Matlab una muestra de* $\mathbf{Y}$ *a partir de la muestra (3.1) obtenida en (a).*

SOLUCIÓN

(a) Puesto que las dos componentes de $\mathbf{X}$ son independientes generamos sendas muestras independientes entre sí y de tamaño n de la normal estándar:

```
n = 100 ;
X = randn(n,2) ;
```

(b) Observemos que

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{A}\mathbf{X}\,,$$

siendo

$$\mathbf{A} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 \\ \sqrt{\sigma_{22}}\rho & \sqrt{\sigma_{22}}\sqrt{1-\rho^2} \end{bmatrix}.$$

Por tanto, como $\mathbf{X}$ sigue una distribución normal, el vector $\mathbf{Y}$ también. Además

$$E(\mathbf{Y}) = \boldsymbol{\mu} + \mathbf{A}\, E(\mathbf{X}) = \boldsymbol{\mu}$$

y

$$\mathit{Var}(\mathbf{Y}) = \mathit{Var}(\mathbf{A}\mathbf{X}) = \mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}\,.$$

(c)

```
mu = [ 2 ; 1 ] ;
sigma_11 = 1 ;  sigma_22 = 1.5 ;
rho = 0.6 ;
A = [sqrt(sigma_11) 0
     sqrt(sigma_22)*rho sqrt(sigma_22)*sqrt(1-rho^2) ] ;
Y = ones(n,1) * mu' + X*A' ;
```

PROBLEMA 3.14

Siguiendo el esquema del Problema 3.13, generaremos muestras de una normal tridimensional. Para ello elíjase un tamaño muestral n y genérese una muestra

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \tag{3.3}$$

de $\mathbf{X} \sim N_3(\mathbf{0}, \mathbf{I})$. *A continuación fijemos la matriz de correlaciones*

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.8 \\ 0.7 & 0.8 & 1 \end{pmatrix}.$$

Decídase cuál es la combinación lineal $\mathbf{Y} = \mathbf{AX}$ *de* $\mathbf{X}$ *tal que* $\boldsymbol{\rho} = Var(\mathbf{Y})$. *A partir de esta matriz* $\mathbf{A}$ *y de la muestra (3.3) generada, calcúlense los valores observados de* $\mathbf{Y}$. *Calcúlese la matriz de correlaciones muestral de* $\mathbf{Y}$, $\mathbf{R}$, *y verifíquese si está próxima o no a la poblacional* $\boldsymbol{\rho}$.

Indicación: *Utilícese la descomposición espectral de la matriz de correlaciones,* $\boldsymbol{\rho} = \mathbf{B}\,\mathbf{D}\,\mathbf{B}'$.

SOLUCIÓN

Observemos que basta tomar $\mathbf{A} = \mathbf{B}\mathbf{D}^{1/2}$. El código que nos piden es

```
n = 10000 ;
X = randn(n,3) ;
rho = [ 1 0.9 0.7 ; 0.9 1 0.8 ; 0.7 0.8 1 ] ;
[B,D] = eig(rho) ;
A = B * sqrt(D) ;
Y = X * A' ;
R = corrcoef(Y) ;
```

PROBLEMA 3.15

Sea $\boldsymbol{\mu}$ un vector $p \times 1$ y $\boldsymbol{\Sigma}$ una matriz $p \times p$ simétrica y definida positiva. Fíjese un valor de p y genérense muestras de tamaño n de una normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ para distintos valores de n. Para cada muestra obténganse el vector de medias muestrales, $\overline{\mathbf{x}}$, y la matriz de covarianzas muestrales, $\mathbf{S}$, y compruébese que a medida que aumenta n, los valores de $\overline{\mathbf{x}}$ y $\mathbf{S}$ se van acercando a $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, respectivamente.

Indicación:
El vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ tiene ley normal p-variante si existen p variables aleatorias independientes con ley $N(0,1)$, $Y_1, Y_2, \ldots, Y_p$, tales que

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{A}\,\mathbf{Y}, \tag{3.4}$$

donde $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)'$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_p)'$ y $\mathbf{A}$ es una matriz $p \times p$. Si las p columnas de $\mathbf{A}$ no son linealmente independientes, alguna de las X_i puede expresarse como combinación lineal de las otras; en caso contrario, se trata de una distribución p-variante no singular.
Si el vector $\mathbf{X}$ verifica (3.4), entonces

$$E(\mathbf{X}) = \boldsymbol{\mu}, \quad Var(\mathbf{X}) = \mathbf{A}'\,\mathbf{A},$$

y se dice que $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde $\boldsymbol{\Sigma} = \mathbf{A}'\,\mathbf{A}$ es definida positiva si $\mathbf{A}$ es regular. Por ejemplo, $\mathbf{A}$ puede ser la matriz de Cholesky de $\boldsymbol{\Sigma}$ (ver Peña 2002), que calculamos en Matlab con la orden `A = chol(Sigma)`.

SOLUCIÓN

Una posible solución del problema consiste en construir una función Matlab que calcule el vector de medias y la matriz de covarianzas para una muestra de tamaño n generada siguiendo la indicación anterior. Para poder llamar a la función dentro de Matlab, ésta debe guardarse en un fichero con el mismo nombre de la función y extensión m, en este caso, `nmult.m`.

```
% funcion [m,S]=nmult(mu,A,n)
%
%  entradas: mu es el vector px1 de medias poblacionales,
%            A  es una matriz cuadrada pxp, de manera que la
%               matriz de covarianzas poblacionales es Sigma=A'A,
%            n  es el tamanio muestral,
%
%  salidas:  m  es el vector de medias muestrales,
%            S  es la matriz de covarianzas muestrales.
%
function [m,S] = nmult(mu,A,n)
% generacion de una muestra p-variante N(0,Id)
[p,p] = size(A);
Y = randn(n,p);
```

```
% generacion de una muestra p-variante N(mu,A'A)
u = ones(n,1);
X = u*mu'+Y*A;
% vector de medias y matriz de covarianzas
m = mean(X);
S = cov(X,1);
```

Dentro de Matlab, y por ejemplo para $\boldsymbol{\mu} = (2, 3, 4)'$, $n = 500, 1000, 5000$ y

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$$

haremos:

```
mu = [2 3 4];
A = [1 -1 1; -1 1 0; 0 1 -1];
[m_500,S_500] = nmult(mu,A,500)
[m_1000,S_1000] = nmult(mu,A,1000)
[m_5000,S_5000] = nmult(mu,A,5000)
```

y compararemos `m_500`, `m_1000`, `m_50000` con $\boldsymbol{\mu}$ y `S_500`, `S_1000`, `S_5000` con $\boldsymbol{\Sigma} = \mathbf{A}'\,\mathbf{A}$, respectivamente.

PROBLEMA 3.16

Una distribución muy relacionada con la ley normal multivariante, y que es el análogo multivariante de la ley χ^2, es la distribución Wishart. Dados $\mathbf{X}_1, \ldots, \mathbf{X}_n$ vectores aleatorios i.i.d.$\sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, la matriz $p \times p$

$$\mathbf{Q} = \sum_{i=1}^{n} \mathbf{X}_i\,\mathbf{X}_i' \sim W_p(\boldsymbol{\Sigma}, n)$$

sigue una ley Wishart con parámetro de escala $\boldsymbol{\Sigma}$ y n grados de libertad.
Dadas las variables aleatorias $\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{I})$ y $\mathbf{Q} \sim W_p(\mathbf{I}, n)$ estocásticamente independientes, la variable aleatoria

$$T^2 = n\,\mathbf{Z}'\,\mathbf{Q}^{-1}\,\mathbf{Z} \sim T^2(p, n)$$

sigue una ley T^2 de Hotelling con p y n grados de libertad. Si $p = 1$, entonces $T^2(1, n)$ es el cuadrado de una variable aleatoria con ley t de Student y n grados de libertad. En general, $T^2(p, n)$ es proporcional a una F de Fisher

$$\frac{n - p + 1}{n\,p}\,T^2(p, n) = F(p, n - p + 1). \tag{3.5}$$

La variable T^2 se utiliza de manera análoga a la ley t de Student, en contrastes sobre medias multivariantes.
Para p y n fijos, genérese una muestra de tamaño N de una ley $T^2(p, n)$ de Hotelling. Represéntense los resultados mediante un histograma.

SOLUCIÓN

Sea $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ una muestra de vectores i.i.d. con distribución $N_p(\mathbf{0}, \mathbf{I})$. Puesto que

$$\overline{\mathbf{x}} \sim N_p\left(\mathbf{0}, \frac{1}{n}\mathbf{I}\right)$$

y

$$n\,\mathbf{S} \sim W_p(\mathbf{I}, n-1),$$

tenemos que

$$(n-1)\,\overline{\mathbf{x}}'\,\mathbf{S}^{-1}\,\overline{\mathbf{x}} \sim T^2(p, n-1).$$

Podemos construir una función Matlab que genere muestras de tamaño N de una ley $T^2(p,n)$ de la siguiente manera:

```
%  funcion randT2
%
%  Esta funcion genera una muestra de tamanio N de una ley
%  T^2 de Hotelling con p y n grados de libertad.
%
   function t=randT2(p,n,N)
%
   n = n+1;
   for i = 1:N
      X = randn(n,p);
      m = mean(X);
      S = cov(X,1);
      t(i,1) = (n-1)*m*inv(S)*m';
   end
%  numero de intervalos en el histograma
   if N<10^4
      k = sqrt(N);
   else
      k = 1+3.22*log10(N);
   end
   int = 0:max(t)/k:max(t);
   hist(t,int)
   h = findobj(gca,'Type','patch');
   set(h,'FaceColor','c','EdgeColor','b')
```

Por ejemplo, para $p = 4, n = 100, N = 1000$, dentro de Matlab utilizaremos la orden

```
t=randT2(4,100,1000)
```

La Figura 3.6 contiene el histograma de frecuencias absolutas.

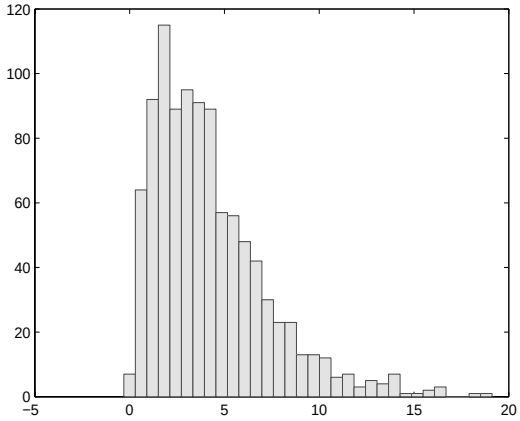


Figura 3.6.
Histograma de una $T^2(4, 100)$ para tamaño muestral $N = 1000$ (Problema 3.16)

PROBLEMA 3.17

Si $\mathbf{A} \sim W_p(\mathbf{\Sigma}, a)$ *y* $\mathbf{B} \sim W_p(\mathbf{\Sigma}, b)$ *son independientes,* $\mathbf{\Sigma}$ *es regular y* $a \geq p$*, la variable aleatoria*

$$\Lambda = \frac{|\mathbf{A}|}{|\mathbf{A} + \mathbf{B}|}$$

tiene una ley Lambda de Wilks , $\Lambda(p, a, b)$*, con parámetros* p*,* a *y* b*.*
La ley Λ *no depende del parámetro* $\mathbf{\Sigma}$ *de* $\mathbf{A}$ *y* $\mathbf{B}$*, por lo que es suficiente considerarla para* $\mathbf{\Sigma} = \mathbf{I}$*. Tiene la misma distribución que un producto de* b *v.a. independientes con distribución Beta, es decir, si* $L \sim \Lambda(p, a, b)$ *entonces*

$$L = \prod_{i=1}^{b} u_i, \quad \textit{donde } u_i \sim \textit{Beta}\left(\frac{a + i - p}{2}, \frac{p}{2}\right).$$

Genérese una muestra de tamaño N *de una ley* Λ *de Wilks. Represéntense los resultados mediante un histograma.*

SOLUCIÓN

Sean

$$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{n_X} \sim N_p(\boldsymbol{\mu}_X, \mathbf{I})$$

e

$$\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{n_Y} \sim N_p(\boldsymbol{\mu}_Y, \mathbf{I})$$

dos muestras independientes de vectores aleatorios i.i.d. Puesto que

$$\mathbf{A} = n_X \, \mathbf{S}_X \sim W_p(\mathbf{I}, n_X - 1)$$

y

$$\mathbf{B} = n_Y \, \mathbf{S}_Y \sim W_p(\mathbf{I}, n_Y - 1)$$

entonces

$$\Lambda = \frac{|\mathbf{A}|}{|\mathbf{A} + \mathbf{B}|} \sim \Lambda(p, n_X - 1, n_Y - 1) \, .$$

Podemos construir una función Matlab que genere muestras de tamaño N de una ley $\Lambda(p, a, b)$, de la siguiente manera

```
%  funcion randWilks
%
%  Esta funcion genera una muestra de tamanio N de una ley
%  Lambda de Wilks con parametros p, a, b. (Atencion: a>=p).
%
   function L = randWilks(p,a,b,N)
   nx = a+1; ny = b+1;
%  los vectores de medias se generan a partir de uniformes, pero
%  tambien podrian introducirse como argumentos de la funcion.
   mux = rand(1,p); muy = 10*rand(1,p);
   ux = ones(nx,1); uy = ones(ny,1);
%
   for i = 1:N
%  generacion de la primera muestra de normales
      Zx = randn(nx,p);
      X = ux*mux+Zx;
      A = nx*cov(X,1);
%  generacion de la segunda muestra de normales
      Zy = randn(ny,p);
      Y = uy*muy+Zy;
      B = ny*cov(Y,1);
%  obtencion de la Lambda de Wilks
      L(i,1) = det(A)/det(A+B);
   end
%  numero de intervalos en el histograma
   if N<10^4
      k = sqrt(N);
   else
      k = 1+3.22*log10(N);
   end
   int = 0:max(L)/k:max(L);
   hist(L,int)
   h = findobj(gca,'Type','patch');
   set(h,'FaceColor','c','EdgeColor','b')
```

Por ejemplo, para $p = 4, a = 19, b = 24$, dentro de Matlab llamaremos a la función

```
L=randWilks(4,19,24,1000)
```

La Figura 3.7 contiene el histograma de frecuencias absolutas.

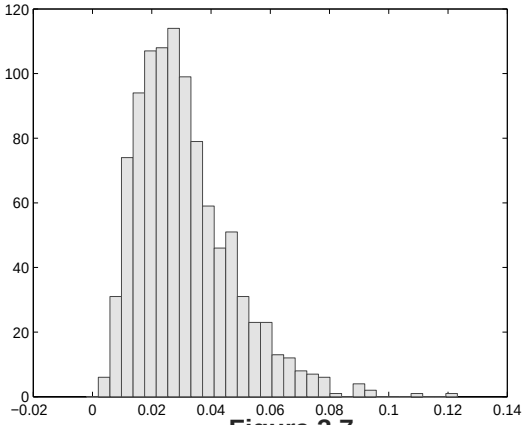


Figura 3.7.
Histograma de una $\Lambda(4, 19, 24)$ para tamaño muestral $N = 1000$ (Problema 3.17)

PROBLEMA 3.18

La Tabla 3.1 contiene las medidas de 5 variables biométricas sobre gorriones hembra, recogidos casi moribundos después de una tormenta. Los primeros 21 sobrevivieron mientras que los 28 restantes no lo consiguieron. Las variables son $X_1 =$ longitud total, $X_2 =$ extensión del ala, $X_3 =$ longitud del pico y de la cabeza, $X_4 =$ longitud del húmero y $X_5 =$ longitud del esternón.
Realícense comparaciones de medias y de covarianzas entre el grupo de supervivientes y el de no supervivientes.

Tabla 3.1.
Medidas biométricas sobre gorriones (Problema 3.18)

Supervivientes					No supervivientes				
X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
156	245	31.6	18.5	20.5	155	240	31.4	18.0	20.7
154	240	30.4	17.9	19.6	156	240	31.5	18.2	20.6
153	240	31.0	18.4	20.6	160	242	32.6	18.8	21.7
153	236	30.9	17.7	20.2	152	232	30.3	17.2	19.8
155	243	31.5	18.6	20.3	160	250	31.7	18.8	22.5
163	247	32.0	19.0	20.9	155	237	31.0	18.5	20.0
157	238	30.9	18.4	20.2	157	245	32.2	19.5	21.4
155	239	32.8	18.6	21.2	165	245	33.1	19.8	22.7
164	248	32.7	19.1	21.1	153	231	30.1	17.3	19.8
158	238	31.0	18.8	22.0	162	239	30.3	18.0	23.1
158	240	31.3	18.6	22.0	162	243	31.6	18.8	21.3
160	244	31.1	18.6	20.5	159	245	31.8	18.5	21.7
161	246	32.3	19.3	21.8	159	247	30.9	18.1	19.0
157	245	32.0	19.1	20.0	155	243	30.9	18.5	21.3
157	235	31.5	18.1	19.8	162	252	31.9	19.1	22.2
156	237	30.9	18.0	20.3	152	230	30.4	17.3	18.6
158	244	31.4	18.5	21.6	159	242	30.8	18.2	20.5
153	238	30.5	18.2	20.9	155	238	31.2	17.9	19.3
155	236	30.3	18.5	20.1	163	249	33.4	19.5	22.8
163	246	32.5	18.6	21.9	163	242	31.0	18.1	20.7
159	236	31.5	18.0	21.5	156	237	31.7	18.2	20.3
					159	238	31.5	18.4	20.3
					161	245	32.1	19.1	20.8
					155	235	30.7	17.7	19.6
					162	247	31.9	19.1	20.4
					153	237	30.6	18.6	20.4
					162	245	32.5	18.5	21.1
					164	248	32.3	18.8	20.9

SOLUCIÓN

Llamamos X e Y a las matrices de datos del grupo de supervivientes y del de no supervivientes, respectivamente. Mediante Matlab calculamos los vectores de medias y las matrices de covarianzas de cada grupo

```
mx=mean(X); my=mean(Y); Sx=cov(X,1); Sy=cov(Y,1);
```

y obtenemos:

```
mx=[157.3810  241.0000   31.4333   18.5000 20.8095]

Sx =[10.5215    8.6667    1.4825    0.8286    1.2249
      8.6667   16.6667    1.8190    1.2476    0.8381
      1.4825    1.8190    0.5060    0.1800    0.2283
      0.8286    1.2476    0.1800    0.1676    0.1262
      1.2249    0.8381    0.2283    0.1262    0.5475]

my=[158.4286  241.5714   31.4786   18.4464   20.8393]

Sy=[14.5306   16.5765    2.1628    1.6837    2.8260
    16.5765   31.3878    3.2765    2.8449    3.9204
     2.1628    3.2765    0.7024    0.4528    0.5391
     1.6837    2.8449    0.4528    0.4189    0.4878
     2.8260    3.9204    0.5391    0.4878    1.2738]
```

Comparación de covarianzas. Supondremos que X es una muestra aleatoria simple de tamaño n_X de una ley normal multivariante $\mathbf{X} \sim N_5(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ y que Y es otra muestra aleatoria simple independiente de la anterior y de tamaño n_Y de una ley normal multivariante $\mathbf{Y} \sim N_5(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$. Queremos contrastar la hipótesis de igualdad de covarianzas, es decir:

$$H_0 : \boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma} \tag{3.6}$$

Utilizaremos el contraste de la razón de verosimilitudes, cuyo estadístico es

$$\lambda_R = \frac{|\mathbf{S}_X|^{n_X/2}\,|\mathbf{S}_Y|^{n_Y/2}}{|\mathbf{S}|^{n/2}},$$

donde $\mathbf{S}_X$ y $\mathbf{S}_Y$ son las matrices de covarianzas muestrales de cada grupo, $n = n_X + n_Y$ y $\mathbf{S}$ es la matriz de covarianzas común, que se obtiene mediante la siguiente ponderación:

$$\mathbf{S} = \frac{n_X\,\mathbf{S}_X + n_Y\,\mathbf{S}_Y}{n_X + n_Y}.$$

Bajo la hipótesis nula dada por (3.6), tenemos que

$$-2\log(\lambda_R) \sim \chi^2_q,$$

donde

$$q = (g-1)p(p+1)/2\,,$$

g es el número de grupos y p es el número de variables.

Para implementar este contraste mediante Matlab y teniendo en cuenta que

$$-2\log(\lambda_R) = n\,\log|\mathbf{S}| - (n_X \log|\mathbf{S}_X| + n_Y \log|\mathbf{S}_Y|). \tag{3.7}$$

escribimos:

```
nx = 21 ;  ny = 28 ; n = nx+ny ;
S = (nx*Sx+ny*Sy)/n ;
logR = n*log(det(S)) - (nx*log(det(Sx))+ny*log(det(Sy)))
percentil = chi2inv(0.95,15)
p_valor = 1-chi2cdf(logR,15)
```

El valor que obtenemos para el estadístico (3.7) es `logR = 12.5322`. Rechazaremos la hipótesis (3.6) si el valor de este estadístico pertenece a la región crítica $[x_{1-\alpha}, +\infty)$, donde $x_{1-\alpha}$ es el percentil $(1-\alpha)\,100\%$ de una χ^2_{15}. Para un nivel de significación $\alpha = 0.05$ la instrucción `chi2inv(0.95,15)` calcula este percentil. También podemos calcular el *p-valor* del contraste mediante `1-chi2cdf(logR,15)` y comparar este valor directamente con el nivel de significación. En ambos casos, concluimos que no podemos rechazar la hipótesis (3.6), y por tanto, consideraremos que las matrices de covarianzas poblacionales son iguales.

Comparación de medias. Suponiendo igualdad de covarianzas, queremos contrastar la hipótesis

$$H_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y. \tag{3.8}$$

Dos posibles formas de resolver este contraste son:

(a) utilizando el estadístico basado en la distribución T^2 de Hotelling o

(b) utilizando el estadístico basado en la distribución Λ de Wilks.

(a) Sean $\overline{\mathbf{x}}$ e $\overline{\mathbf{y}}$ los vectores de medias muestrales. El estadístico siguiente

$$T^2 = \frac{n_X\, n_Y}{n_X + n_Y}\,(\overline{\mathbf{x}} - \overline{\mathbf{y}})\,\mathbf{S}^{-1}\,(\overline{\mathbf{x}} - \overline{\mathbf{y}})'$$

tiene una ley T^2 de Hotelling $T^2(p, n_X + n_Y - 2)$. La relación (3.5) entre las leyes T^2 de Hotelling y F de Fisher asegura que el estadístico

$$F = \frac{n_X + n_Y - p - 1}{(n_X + n_Y - 2)\,p}\,T^2$$

sigue una distribución $F(p, n_X + n_Y - p - 1)$. Si llamamos `T2` y `F` a estos estadísticos, resolvemos el contraste mediante:

```
T2 = nx*ny/n*(mx-my)*inv(S)*(mx-my)';
F = (nx+ny-p-1)/((nx+ny)*p)*T2
percentil = finv(0.95,p,nx+ny-1)
p_valor = 1-fcdf(F,p,nx+ny-p-1)
```

Puesto que, para un nivel de significación $\alpha = 0.05$, `F=0.5167` no está contenido en la región crítica $[2.4085, +\infty)$, concluimos que no se puede rechazar la hipótesis (3.8). De manera equivalente, observando el p-valor asociado a este valor de `F`, `p-valor=0.7622`, también concluimos que no existen diferencias significativas entre el grupo de supervivientes y el de no supervivientes.

(b) Consideremos el estadístico siguiente

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = \frac{|\mathbf{W}|}{|\mathbf{T}|},$$

donde $\mathbf{W} = n_X\,\mathbf{S}_X + n_Y\,\mathbf{S}_Y$ es la matriz de dispersión dentro de los grupos (*within*),

$$\mathbf{B} = n_X\,(\overline{\mathbf{x}}-\overline{\mathbf{z}})'(\overline{\mathbf{x}}-\overline{\mathbf{z}}) + n_Y\,(\overline{\mathbf{y}}-\overline{\mathbf{z}})'(\overline{\mathbf{y}}-\overline{\mathbf{z}})$$

es la matriz de dispersión entre los grupos (*between*), $\mathbf{T} = \mathbf{W} + \mathbf{B}$ es la matriz de dispersión total y $\overline{\mathbf{z}} = (n_X\,\overline{\mathbf{x}} + n_Y\,\overline{\mathbf{y}})/n$ es el vector de medias global.
Bajo la hipótesis nula (3.8) el estadístico Λ sigue una ley Lambda de Wilks

$$\Lambda(p, n-g, g-1),$$

siendo g el número de grupos. La aproximación asintótica de Rao da una equivalencia asintótica entre la distribución Λ de Wilks y la ley F de Fisher. La función Matlab `wilkstof.m` calcula esta aproximación.

```
%  funcion wilkstof
%
%  Esta funcion calcula la aproximacion asintotica de Rao
%  de la distribucion Lambda de Wilks, L(p,a,b),
%  hacia la distribucion F(m,n).
%
%  [F,m,n]=wilkstof(L,p,a,b)
%
%  entradas: L es el valor de L(p,a,b)
%            p, a, b son los grados de libertad
%
%  salidas:  F es el valor de la F(m,n)
%            m, n son los grados de libertad
%
   function [F,m,n] = wilkstof(L,p,a,b)
   alpha = a+b-(p+b+1)/2;
   beta = sqrt((p^2*b^2-4)/(p^2+b^2-5));
   gamma = (p*b-2)/4;
   m = p*b;
   n = alpha*beta-2*gamma;
%  se redondea n al entero mas proximo
   if n-floor(n)<0.5
      n = floor(n);
   else
      n = floor(n)+1;
   end
   F = (1-L^(1/beta))/(L^(1/beta))*n/m;
```

Implementemos este segundo contraste mediante Matlab. Empezamos calculando el vector de medias global y las matrices de dispersión dentro de los grupos, entre grupos y total:

```
mz = (nx*mx+ny*my)/n;
W = nx*Sx+ny*Sy;
B = nx*(mx-mz)'*(mx-mz)+ny*(my-mz)'*(my-mz);
T = W+B;
```

El estadístico Λ de Wilks se obtiene haciendo:

```
Lambda = det(W)/det(T)
[F,m,n] = wilkstof(Lambda,5,47,1)
percentil = finv(0.95,m,n)
p_valor = 1-fcdf(F,m,n)
```

Los valores obtenidos son

```
Lambda = 0.9433, F = 0.5167 (m=5,n=43),
percentil = 2.4322, p_valor = 0.7622
```

Dado que el valor de `F` no está contenido en la región crítica $[2.4322, +\infty)$, no podemos rechazar la hipótesis nula de igualdad de medias.

PROBLEMA 3.19

En una fábrica de zumos se diseña el siguiente procedimiento de control de calidad. Se toma una muestra piloto (véase la Tabla 3.2) de $n = 50$ extracciones de zumo cuando el proceso de fabricación funciona correctamente y en ella se mide la concentración de $p = 11$ aminoácidos, $\mathbf{X} = (X_1, \ldots, X_{11})'$. Supóngase que $\mathbf{X}$ sigue una distribución normal. A continuación cada día se observan estas mismas variables con objeto de detectar algún cambio significativo en la calidad del proceso (véase Tabla 3.3). Supóngase que estas sucesivas observaciones, $\mathbf{y}_i$, $i = 1, \ldots, 10$, son independientes de la muestra piloto y entre sí.

Constrúyase un gráfico de control para estos nuevos diez días como se indica a continuación. En primer lugar calcúlense la media $\bar{\mathbf{x}}$ y la matriz de covarianzas $\mathbf{S}$ para la muestra piloto. A continuación para la observación $\mathbf{y}_i$ constrúyase el estadístico

$$T^2(i) = \frac{n}{n+1}(\mathbf{y}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{y}_i - \bar{\mathbf{x}})$$

que debería seguir una $T^2(p, n-1)$ si la distribución de $\mathbf{y}_i$ es la misma que la de la muestra piloto.

Represéntense secuencialmente los valores de $T^2(i)$ en un gráfico y márquese en él un límite de control $LC = \frac{(n-1)\,p}{n-p} F^{\alpha}(p, n-p)$, siendo α el nivel de significación que deseemos fijar ($\alpha = 0.05$, por ejemplo). Párese el proceso de fabricación el primer día i que una observación $\mathbf{y}_i$ esté fuera de la región de control, es decir, $\mathbf{y}_i > LC$.

Tabla 3.2.
Concentraciones de 11 aminoácidos en 50 zumos (Problema 3.19)

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
0.480	5.234	2.620	2.857	0.803	13.897	0.326	0.902	0.164	0.183	4.155
0.245	1.312	2.115	8.077	0.974	9.227	0.252	2.703	-0.006	-0.061	1.995
0.276	3.402	2.527	5.447	0.957	13.474	0.299	2.341	0.094	0.113	3.541
0.482	6.554	2.631	5.134	0.671	12.333	0.259	1.473	0.216	0.112	3.941
0.400	4.011	2.528	3.716	0.805	10.382	0.266	0.697	0.201	0.159	4.361
0.336	4.001	3.083	4.626	0.904	7.834	0.156	0.898	0.130	0.061	2.444
0.379	3.366	2.099	6.142	0.977	17.366	0.384	2.451	0.204	0.063	3.177
0.369	4.550	2.242	3.609	0.672	12.353	0.291	0.975	0.158	0.201	3.185
0.396	5.479	2.231	4.264	0.786	15.248	0.244	1.318	0.064	0.116	3.989
0.325	3.573	2.446	5.087	0.708	10.791	0.183	1.500	0.075	0.122	3.675
0.404	4.195	3.226	4.959	0.948	14.880	0.460	0.910	0.151	0.280	5.071
0.367	4.756	2.891	4.264	0.799	13.443	0.270	0.927	0.195	0.194	3.932
0.340	3.640	3.075	4.937	0.821	13.782	0.296	1.659	0.214	0.107	3.507
0.281	2.872	2.299	4.543	0.926	8.921	0.205	0.901	0.072	0.102	2.567
0.373	4.212	2.769	5.014	1.060	15.577	0.288	1.664	0.175	0.095	3.788
0.356	3.629	3.435	4.694	0.843	11.503	0.253	1.249	0.106	0.198	3.147
0.426	5.087	2.797	3.029	0.758	11.412	0.311	0.912	0.175	0.154	3.759
0.262	2.722	3.439	6.223	1.018	8.324	0.233	1.200	0.083	0.108	3.065
0.422	5.769	1.948	4.525	0.576	15.151	0.342	1.282	0.014	0.087	4.773
0.242	2.074	3.090	6.822	0.987	10.655	0.274	1.858	0.065	0.072	2.754
0.288	3.413	3.338	5.562	1.054	9.265	0.276	1.830	0.181	0.071	2.710
0.409	4.701	3.340	5.531	1.237	13.800	0.274	1.598	0.159	0.102	3.032
0.382	4.362	2.588	3.941	0.779	14.441	0.265	1.480	0.213	0.147	3.372
0.277	3.261	2.730	4.335	0.747	7.909	0.181	1.014	0.102	0.108	2.910
0.416	3.511	2.822	5.128	0.992	15.695	0.298	1.864	0.268	0.108	4.097
0.238	2.840	3.180	6.392	1.293	9.059	0.209	1.529	0.120	0.043	3.000
0.544	6.523	3.333	3.431	0.759	13.712	0.334	0.423	0.128	0.240	5.209
0.404	4.119	2.689	4.599	0.744	13.960	0.264	1.241	0.099	0.126	4.185
0.384	4.126	2.440	5.626	0.965	11.960	0.224	1.647	0.203	0.086	3.102
0.290	2.823	2.731	6.063	0.688	7.677	0.217	1.343	0.065	0.073	3.250
0.598	5.807	2.525	4.633	0.889	16.131	0.368	1.462	0.221	0.169	4.544
0.337	4.067	2.902	4.826	0.772	14.203	0.343	1.577	0.167	0.074	3.355
0.403	4.327	2.660	4.993	0.863	14.668	0.402	1.720	0.125	0.091	3.617
0.241	4.281	2.984	4.369	0.828	9.670	0.243	1.036	0.201	0.105	3.089
0.412	4.038	3.731	4.341	0.971	12.550	0.244	1.197	0.135	0.180	3.309
0.154	1.840	3.533	6.902	1.308	8.954	0.190	2.047	0.091	0.018	1.608
0.352	5.170	2.945	2.187	0.866	11.566	0.306	0.765	0.194	0.165	2.959
0.288	3.336	3.430	5.054	0.896	10.608	0.258	1.017	0.104	0.175	2.689
0.447	5.060	3.240	5.462	0.937	18.099	0.339	1.762	0.196	0.164	3.649
0.420	5.828	2.898	4.121	0.793	14.167	0.347	1.133	0.180	0.199	4.181
0.492	5.230	2.116	3.516	0.584	16.289	0.374	1.241	0.262	0.188	4.687
0.385	4.707	2.350	4.655	0.882	15.452	0.357	1.789	0.208	0.153	3.213
0.354	4.626	2.854	4.885	0.753	14.250	0.273	1.332	0.072	0.098	3.228
0.244	3.112	3.245	6.687	1.095	11.960	0.240	2.001	0.177	0.080	2.440
0.221	2.715	2.848	5.216	0.978	6.625	0.137	1.202	0.075	0.015	1.833
0.374	2.819	2.694	5.560	0.804	10.830	0.268	1.472	0.069	0.137	2.838
0.416	3.943	2.908	6.660	1.076	14.812	0.313	2.033	0.173	0.069	3.716
0.356	3.874	2.739	4.778	0.894	11.158	0.215	1.099	0.149	0.093	3.510
0.410	4.898	2.362	3.565	0.630	11.763	0.342	0.783	0.119	0.169	4.037
0.246	2.761	2.914	4.860	0.799	5.649	0.168	1.192	0.016	0.069	2.180

Tabla 3.3.
Concentraciones de aminoácidos en 10 nuevos zumos (Problema 3.19)

Día	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}
1	0.275	3.693	2.785	6.812	1.086	12.206	0.262	2.152	0.091	0.106	2.851
2	0.295	3.401	2.594	5.903	0.964	9.945	0.189	1.719	0.069	0.058	2.271
3	0.370	3.865	2.935	7.034	1.122	18.572	0.354	2.354	0.148	0.043	3.779
4	0.385	3.585	3.601	5.454	1.139	11.033	0.255	0.857	0.078	0.130	3.625
5	0.248	3.188	2.966	7.090	1.205	7.800	0.199	1.657	0.046	0.024	2.733
6	0.480	4.512	2.142	4.533	0.762	18.385	0.345	1.710	0.093	0.167	4.872
7	0.417	5.260	2.554	3.404	0.773	13.679	0.277	0.908	0.122	0.161	3.734
8	0.327	4.388	3.110	4.396	0.774	9.041	0.213	0.669	0.129	0.141	3.725
9	0.251	3.125	2.589	6.390	1.106	13.410	0.235	1.898	0.107	0.044	2.864
10	0.422	4.810	2.002	3.322	1.144	15.986	0.348	1.147	0.154	0.178	3.511

SOLUCIÓN

Suponemos que ya hemos introducido en Matlab las matrices de datos `X` e `Y` que contienen la muestra piloto y las nuevas observaciones respectivamente. Con el siguiente código conseguimos el gráfico de control de la Figura 3.8 que nos indica que el proceso de producción está fuera de control en el día 10.

```
m = mean(X) ; S = cov(X,1) ;
[NumDias,p] = size(Y) ;
alpha = 0.05 ;
LC = ((n-1)*p/(n-p)) * finv(1-alpha,p,n-p) ;
T_i = 0;
T = [ ] ;
i = 1 ;
while (T_i <= LC) & (i <= NumDias)
    T_i = n*(Y(i,:)-m)*inv(S)*(Y(i,:)-m)'/(n+1) ;
    T = [T ; T_i] ;
    i = i+1 ;
end
plot([1:i-1]',T,'ko-','MarkerFaceColor','k')
hold on
plot([1:i-1],LC*ones(1,i-1),'k--','LineWidth',1.5)
hold on
text(1.5,LC+1.2,'UCL','FontSize',14)
if (i <= NumDias) | (T_i>UCL)
   plot(i-1,T_i,'ko','MarkerSize',12)
end
xlabel('Dia')
ylabel('T_i^2')
```

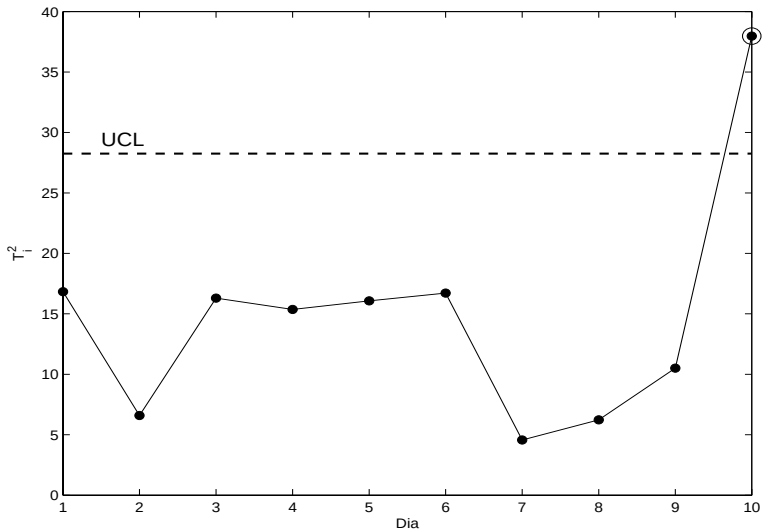


Figura 3.8.
Gráfico de control para datos de zumos (Problema 3.19)

CAPÍTULO 4

Análisis de componentes principales

El problema de reducción de la dimensión subyace tras la mayoría de los métodos de Análisis Multivariante. Genéricamente puede plantearse de la manera siguiente: ¿Es posible describir la información contenida en unos datos mediante un número de variables menor que el de variables observadas?
El análisis de componentes principales parte de una matriz de datos (centrada) de n filas y p columnas, que puede considerarse como una muestra de tamaño n de un vector aleatorio de dimensión p,

$$\mathbf{X} = (X_1, \ldots, X_p)'.$$

Se considera una combinación lineal (univariante) de $\mathbf{X}$,

$$\mathbf{y} = \mathbf{X}'\mathbf{t}\,,$$

donde $\mathbf{t}$ es un vector de pesos de dimensión p. La primera componente principal aparece como solución al problema de encontrar el vector $\mathbf{t}$ que maximiza la varianza de $\mathbf{Y}$ con la condición de normalización $\mathbf{t}'\mathbf{t} = 1$. En otras palabras, la expresión $var(\mathbf{Y})$ en función del vector de pesos $\mathbf{t}$ da lugar a un problema variacional que tiene por solución la primera componente principal. Este problema equivale a encontrar los autovalores y autovectores de la matriz de covarianzas de $\mathbf{X}$. De manera que las sucesivas componentes principales se obtienen de la diagonalización de la matriz de covarianzas de $\mathbf{X}$,

$$\mathbf{S} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}'\,,$$

donde $\mathbf{T}$ es una matriz ortogonal $p \times p$ cuyas columnas son los coeficientes de las componentes principales.

PROBLEMA 4.1

Sea la matriz de varianzas-covarianzas poblacionales

$$\mathbf{\Sigma} = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 5 \end{pmatrix}$$

correspondiente a un vector aleatorio $\mathbf{X} = (X_1, X_2, X_3)'$ *de media cero.*

(a) *Calcúlense los autovalores y autovectores de* $\mathbf{\Sigma}$.

(b) *Escríbase el vector* $\mathbf{Y} = (Y_1, Y_2, Y_3)'$ *de componentes principales e indíquese qué proporción de la varianza total explica cada componente.*

(c) *Represéntese el vector* $\mathbf{X}$ *original en el plano de las dos primeras componentes principales. Concrétese esta representación para la observación* $\mathbf{x} = (2, 2, 1)'$.

SOLUCIÓN

(a) Los autovalores de $\mathbf{\Sigma}$, ordenados de mayor a menor, son $\lambda_1 = 6$, $\lambda_2 = 3$ y $\lambda_3 = 2$. Los correspondientes autovectores normalizados son $\mathbf{e}_1 = (1, 1, 2)'/\sqrt{6}$, $\mathbf{e}_2 = (1, 1, -1)'/\sqrt{3}$ y $\mathbf{e}_3 = (1, -1, 0)'/\sqrt{2}$.

(b) Las componentes principales son

$$\begin{aligned} Y_1 &= \mathbf{e}_1'\mathbf{X} = \frac{1}{\sqrt{6}}(X_1 + X_2 + 2X_3) \\ Y_2 &= \mathbf{e}_2'\mathbf{X} = \frac{1}{\sqrt{3}}(X_1 + X_2 - X_3) \\ Y_3 &= \mathbf{e}_3'\mathbf{X} = \frac{1}{\sqrt{2}}(X_1 - X_2). \end{aligned}$$

La varianza total es

$$VT(\mathbf{\Sigma}) = tr(\mathbf{\Sigma}) = 11\,.$$

La proporción de $VT(\mathbf{\Sigma})$ explicada por la primera componente principal es

$$\frac{var(Y_1)}{VT(\mathbf{\Sigma})} = \frac{\lambda_1}{11} \simeq 54.5\%.$$

Análogamente la explicada por Y_2 e Y_3 es 27.3% y 18.2% respectivamente.

(c) Para expresar $\mathbf{X}$ en el plano de Y_1 e Y_2 debe realizarse el producto escalar de $\mathbf{X}$ por las direcciones dadas por $\mathbf{e}_1$ y $\mathbf{e}_2$. Para $\mathbf{x}$ el resultado es el punto $(y_1, y_2) = (\sqrt{6}, \sqrt{3})$.

PROBLEMA 4.2

Dados los datos de la Tabla 2.1, considérense únicamente las variables $X_1 =$ duración de la hipoteca y $X_2 =$ precio y denótese por $\mathbf{X}$ el vector $(X_1, X_2)'$.

(a) *Calcúlense el vector $\bar{\mathbf{x}}$ y la matriz de covarianzas muestral $\mathbf{S}$.*

(b) *Determínense las componentes principales muestrales Y_1 e Y_2 y sus varianzas.*

(c) *Hállese la proporción de varianza explicada por Y_1.*

(d) *Calcúlense los coeficientes de correlación corr(Y_1, X_k), para $k = 1, 2$. Interprétese la primera componente principal.*

SOLUCIÓN

(a) La media es $\bar{\mathbf{x}} = (19.05, 1.57)'$ y la matriz de covarianzas es

$$\mathbf{S} = \begin{pmatrix} 56.97 & 5.17 \\ & 0.89 \end{pmatrix}.$$

(b) Los autovalores de $\mathbf{S}$ son $\lambda_1 = 57.44$ y $\lambda_2 = 0.42$. Los correspondientes autovectores normalizados son $\mathbf{e}_1 \simeq (0.99, 0.09)'$ y $\mathbf{e}_2 \simeq (0.09, -0.99)'$. Por tanto, las componentes principales de $\mathbf{S}$ tienen la expresión

$$\begin{aligned} Y_1 &= \mathbf{e}_1'(\mathbf{X} - \bar{\mathbf{X}}) = 0.99(X_1 - 19.05) + 0.09(X_2 - 0.42) \\ Y_2 &= \mathbf{e}_2'(\mathbf{X} - \bar{\mathbf{X}}) = 0.09(X_1 - 19.05) - 0.99(X_2 - 0.42). \end{aligned}$$

La varianza de una componente principal es el autovalor de $\mathbf{S}$ que la determina, luego

$$var(Y_1) = \lambda_1 = 57.44 \text{ y } var(Y_2) = \lambda_2 = 0.42\,.$$

(c) La proporción de $VT(\mathbf{S})$ explicada por Y_1 es $var(Y_1)/VT(\mathbf{S}) \simeq 99\%$.

(d) Las correlaciones entre la primera componente y las variables X_i son

$$corr(Y_1, X_1) = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{s_{11}}} = \frac{0.99\sqrt{57.44}}{\sqrt{56.97}} \simeq 0.99,$$

y

$$corr(Y_1, X_2) = \frac{e_{12}\sqrt{\lambda_1}}{\sqrt{s_{22}}} = 0.72.$$

El hecho de que la primera componente principal (que es esencialmente X_1) explique gran parte de la variabilidad del sistema es debido a que la varianza muestral de X_1 es mucho mayor que la de X_2 y eso hace que la varianza sea considerablemente mayor a lo largo de la dirección dada por el vector $\mathbf{e}_1$. En este caso conviene estandarizar los datos y realizar un nuevo análisis de componentes principales sobre la matriz resultante. Esto equivale a obtener las componentes principales a partir de la matriz de correlaciones.

PROBLEMA 4.3

Calcúlese la matriz de correlaciones $\mathbf{R}$ *asociada a la matriz* $\mathbf{S}$ *del Problema 4.2.*

(a) *Determínense las componentes principales a partir de* $\mathbf{R}$ *y sus varianzas.*

(b) *Hállese la proporción de varianza explicada por la primera componente.*

(c) *Calcúlense los coeficientes de correlación entre la primera componente y las variables* X_i *estandarizadas.*

(d) *Compárense las componentes principales obtenidas en (a) con las componentes obtenidas en el apartado (b) del ejercicio anterior. ¿Qué es más adecuado: determinar las componentes principales a partir de* $\mathbf{R}$ *o de* $\mathbf{S}$*?*

SOLUCIÓN

(a) La matriz

$$\mathbf{R} = \begin{pmatrix} 1 & 0.72 \\ 0.72 & 1 \end{pmatrix}.$$

tiene autovalores $\lambda_1 = 1.72$ y $\lambda_2 = 0.28$ y autovectores

$$\mathbf{e}_1 = (0.71, 0.71)' \text{ y } \mathbf{e}_2 = (-0.71, 0.71)'.$$

Por tanto, las componentes principales de $\mathbf{R}$ son

$$Y_1 = \mathbf{e}_1'\mathbf{Z} = 0.71Z_1 + 0.71Z_2 \quad \text{e} \quad Y_2 = \mathbf{e}_2'\mathbf{Z} = -0.71Z_1 + 0.71Z_2,$$

donde $Z_1 = (X_i - 19.05)/7.55$, $Z_2 = (X_2 - 1.57)/0.94$ y $\mathbf{Z} = (Z_1, Z_2)'$ es el vector $\mathbf{X}$ estandarizado.

(b) La variabilidad total viene medida por $VT(\mathbf{R}) = tr(\mathbf{R}) = 2$ y la proporción de la misma explicada por Y_1 es $\lambda_1 / VT(\mathbf{R}) = 1.72/2 = 86\%$.

(c) Los coeficientes de correlación entre Y_1 y las variables Z_i son:

$$corr(Y_1, Z_1) = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{r_{11}}} = 0.93\,, \quad corr(Y_1, Z_2) = 0.93.$$

(d) La primera componente principal de $\mathbf{R}$ da ahora igual peso mayor a las variables X_1 y X_2. Según lo comentado en el Problema 4.2 sería más adecuado calcular las componentes principales a partir de $\mathbf{R}$.

PROBLEMA 4.4

La Tabla 4.1 contiene 11 indicadores económicos y sociales de 96 países. Las variables observadas son: X_1 = *Tasa anual de crecimiento de la población,* X_2 = *Tasa de mortalidad infantil por cada 1000 nacidos vivos,* X_3 = *Porcentaje de mujeres en la población activa,* X_4 = *PNB en 1995 (en millones de dólares),* X_5 = *Producción de electricidad (en millones* kW/h*),* X_6 = *Líneas telefónicas por cada 1000 habitantes,* X_7 = *Consumo de agua per cápita,* X_8 = *Proporción de la superficie del país cubierta por bosques,* X_9 = *Proporción de deforestación anual,* X_{10} = *Consumo de energía per cápita,* X_{11} = *Emisión de CO2 per cápita. Realícese un análisis de componentes principales y razónese a partir de qué matriz ,* $\mathbf{S}$ *o* $\mathbf{R}$*, es más adecuado. Interprétense las dos primeras componentes.*

SOLUCIÓN

Observemos primero que las unidades de medida de las variables X_i son muy distintas (porcentajes, dólares, kWh, ...). Además, las elevadas varianzas de X_4 y X_5 hacen prever que un análisis de componentes principales realizado a partir de la matriz de covarianzas $\mathbf{S}$ dará como resultado una primera y segunda componentes principales que coincidirán básicamente con estas dos variables observadas. Por tanto, el análisis de componentes principales debe llevarse a cabo a partir de la matriz de correlaciones $\mathbf{R}$. Esto equivale a estandarizar cada una de las X_i a media cero y varianza unidad y considerar la matriz de covarianzas de las variables estandarizadas. La siguiente función Matlab realiza el análisis de componentes principales, primero a partir de $\mathbf{S}$ y, en segundo lugar, a partir $\mathbf{R}$.

```
% COMP
%
% La funcion [T1,Y1,acum1,T2,Y2,acum2]=comp(X) calcula las
% componentes principales de una matriz de datos X (n,p).
% Devuelve :
%   T1 componentes principales a partir de la matriz S
%   Y1 representacion de los datos
%   acum1 porcentajes acumulados
%   T2 componentes principales a partir de la matriz R
%   Y2 representacion de los datos
%   acum2 porcentajes acumulados
%
 function [T1,Y1,acum1,T2,Y2,acum2]=comp(X)
 [n,p] = size(X);
% Vector de etiquetas para los individuos.
 for i = 1:n
   lab(i,:)=sprintf('%3g',i);
 end
% Matriz de centrado y matriz de datos centrados.
 H = eye(n)-ones(n)/n;
 X = H*X;
% Calculo de las matrices de covarianzas y de correlaciones.
 S = cov(X,1); R = corr(X);
%
```

Tabla 4.1.

Indicadores económicos y sociales sobre países del mundo (Problema 4.4)

País	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1. Albania	1	30	41	2199	3903	12	94	53	0	341	1.2
2. Angola	3	124	46	4422	955	6	57	19	0.7	89	0.5
3. Arabia Saudi	4.3	21	13	133540	91019	96	497	1	0	4566	13.1
4. Argelia	2.5	34	24	44609	19883	42	180	2	0.8	906	3
5. Argentina	1.3	22	31	278431	65962	160	1043	22	0.1	1504	3.5
6. Australia	1.4	6	43	337909	167155	510	933	19	0	5341	15.3
7. Austria	0.6	6	41	216547	53259	465	304	47	-0.4	3301	7.2
8. Bangladesh	2	79	42	28599	9891	2	220	6	4.1	64	0.2
9. Bélgica	0.3	8	40	250710	72236	457	917	20	-0.3	5120	10.1
10. Benin	3	95	48	2034	6	5	26	45	1.3	20	0.1
11. Bielorrusia	0.4	13	49	21356	31397	190	295	31	-0.4	2392	9.9
12. Bolivia	2.3	69	37	5905	2824	35	201	45	1.2	373	1
13. Brasil	1.6	44	35	579787	260682	75	246	66	0.6	718	1.4
14. Bulgaria	-0.6	15	48	11225	381333	335	1544	33	-0.2	2438	6.4
15. Camerún	2.9	56	38	8615	2740	4	38	44	0.6	103	0.2
16. Canadá	1.3	6	45	573695	554227	590	1602	49	-1.1	7854	14.4
17. Colombia	1.8	26	37	70263	43354	100	174	52	0.7	622	1.8
18. Congo	3.1	90	43	1784	435	8	20	58	0.2	331	1.6
19. Corea del Norte	1.8	26	45	12870	38000	47	687	74	0	1129	11.2
20. Corea del Sur	0.9	10	40	435137	164993	415	632	66	0.1	2982	6.6
21. Costa de Marfil	3.4	86	33	9248	2305	8	66	34	1	103	0.5
22. Costa Rica	2.5	13	30	8884	4772	164	780	28	3	558	1.2
23. Cuba	0.9	9	38	7150	10982	32	870	16	1	923	2.6
24. Chile	1.6	12	32	59151	25276	132	1626	12	-0.1	1012	2.6
25. China	1.3	34	45	744890	928083	34	461	13	0.7	664	2.3
26. Dinamarca	0.2	6	46	156027	40097	613	233	12	0	3977	10.4
27. R. Dominicana	2	37	29	11390	6182	79	446	22	2.9	337	1.4
28. Ecuador	2.3	36	26	15997	8256	61	581	43	1.8	565	1.8
29. Egipto	2.2	56	29	45507	51947	46	956	0	0	600	1.5
30. El Salvador	1.8	36	34	9057	3211	53	245	6	2.3	370	0.7
31. E. Árabes Unidos	5.8	16	13	42806	18870	283	884	0	0	10531	33.9
32. Eslovaquia	0.3	11	48	15848	24740	208	337	38	0.1	3243	7
33. España	0.2	7	36	532347	161654	385	781	51	0	2458	5.7
34. Etiopía	2.6	112	41	5722	1293	2	51	13	0.3	22	0.1
35. Filipinas	2.3	39	37	71865	27062	21	686	26	3.4	316	0.8
36. Finlandia	0.4	5	48	105174	65546	550	440	77	0	5997	8.2
37. Francia	0.5	6	44	1451051	476200	558	665	25	-0.1	4042	6.3
38. Gabón	2.9	89	44	3759	933	30	57	71	0.6	652	5.5
39. Ghana	3	73	51	6719	6115	4	35	42	1.4	93	0.2
40. Grecia	0.5	8	36	85885	40623	493	523	47	0	2260	7.2
41. Guatemala	2.9	44	26	14255	3161	27	139	39	1.8	210	0.6
42. Haití	2	72	43	1777	362	8	7	1	5.1	29	0.1
43. Países Bajos	0.6	6	40	371039	79647	525	518	10	-0.3	4580	9.2
44. Honduras	3	45	30	3566	2672	29	294	41	2.2	204	0.6
45. Hungría	-0.3	11	44	42129	33486	185	661	18	-0.5	2383	5.8
46. India	1.9	68	32	319660	386500	13	612	17	0.6	248	0.9
51. Indonesia	1.7	51	40	190105	53414	17	96	60	1.1	366	1
47. Irak	2.7	108	18	24600	27060	33	4575	4	0.1	1213	3.4
48. Irán	3.2	45	24	113400	79128	79	1362	11	0	1505	4
49. Irlanda	0.1	6	33	52765	17105	365	233	6	-1.2	3137	8.7
50. Islandia	1.1	4	44	6686	4780	555	636	1	0	7932	6.8
52. Israel	2.7	8	40	87875	32781	418	408	6	-0.3	2717	8.1
53. Jordania	4.7	31	21	6354	5076	73	173	1	-1	1067	3
54. Kenia	2.9	58	46	7583	3539	9	87	2	0.6	110	0.2
55. Kuwait	-0.3	11	28	28941	22798	230	525	0	0	8622	11.2
56. Líbano	2.3	32	28	10673	5184	82	271	8	0.6	964	2.9
57. Libia	3.6	61	21	23400	17800	59	880	0	-1.4	2499	8.1
58. Malasia	2.5	12	37	78321	39093	166	768	54	2.1	1699	3.8
59. Marruecos	2	55	35	29545	11100	43	427	20	-1.4	327	1.1
60. México	2.1	33	31	304596	147926	96	899	25	1.3	1561	3.8
61. Mozambique	1.8	113	48	1353	490	3	55	22	0.8	40	0.1
62. Birmania	1.8	83	43	35840	3500	3	101	44	1.3	49	0.1
63. Nepal	2.5	91	40	4391	927	4	150	37	1	28	0.1
64. Nicaragua	3.1	46	36	1659	1688	23	367	50	1.9	300	0.6
65. Nigeria	2.9	80	36	28411	15530	4	41	17	0.7	162	0.9
66. Noruega	0.5	5	46	136077	113488	556	488	31	-1.4	5318	14.1
67. Nueva Zelanda	1	7	44	51655	35135	479	589	28	0	4245	7.6
68. Omán	4.5	18	15	10578	6187	77	564	19	0	2392	5.3
69. Pakistán	3	90	26	59991	58529	16	2053	2	3.5	254	0.6
70. Panamá	1.9	23	34	7253	3380	114	754	42	1.9	618	1.7
71. Paraguay	2.7	41	29	8158	36415	31	109	32	2.8	299	0.6
72. Perú	2.1	47	29	55019	15563	47	300	53	0.4	367	1
73. Polonia	0.4	14	46	107829	135347	148	321	28	-0.1	2401	8.9
74. Portugal	-0.1	7	43	96829	31380	361	739	34	-0.5	1827	4.8
75. Reino Unido	0.3	6	43	1094734	325383	502	205	10	-1.1	3732	9.8
76. Rep. Checa	0	8	47	39990	58705	236	266	34	0	3868	13.1
77. Rumanía	0	23	44	33488	55136	131	1134	27	0	1733	5.4
78. Senegal	2.8	62	42	5070	1002	10	202	39	0.7	97	0.4
79. Singapur	1.8	4	38	79831	20046	478	84	7	2.3	8103	17.7
80. Siria	3.1	32	26	15780	15186	63	435	4	-4.3	997	3.3
81. Sri Lanka	1.3	16	35	12616	4387	11	503	27	1.4	97	0.3
82. Sudán	2.2	77	28	7510	1333	3	633	18	1.1	66	0.1
83. Suecia	0.6	4	48	209720	142895	681	341	68	0	5723	6.6
84. Suiza	0.8	6	40	286014	65724	613	173	30	-0.6	3629	6.4
85. Suráfrica	2.3	50	37	130918	189316	95	359	4	-0.8	2146	7.5
86. Tailandia	1.3	35	46	159630	71177	59	602	25	3.5	769	2
87. Tanzania	3.1	82	49	3703	1913	3	40	38	1.2	34	0.1
88. Túnez	2.1	39	30	16369	6714	58	381	4	-1.9	595	1.6
89. Turquia	1.9	48	35	169452	78322	212	585	26	0	957	2.5
90. Ucrania	0.1	15	49	84084	202995	157	673	16	-0.3	3180	11.7
91. Uruguay	0.6	18	40	16458	7617	196	241	4	-0.6	629	1.6
92. Venezuela	2.4	23	33	65382	73116	111	382	52	1.2	2186	5.7
93. Vietnam	2.2	41	49	17634	12270	11	414	26	1.5	101	0.3
94. Yemen	4.2	100	29	4044	2159	12	335	8	0	206	0.7
95. Zambia	2.6	109	45	3605	7785	8	186	43	1.1	149	0.3
96. Zimbabue	2.8	55	44	5933	7334	14	136	23	0.7	438	1.8

```
% Componentes principales a partir de la matriz de covarianzas.
% Ordenacion de los valores propios segun la variabilidad
% explicada (de mayor a menor). D1 es un vector fila.
% Las filas de T1 son los vectores propios ordenados.
%
 [T1,D1] = eigsort(S); T1 = T1';
% Corregimos los signos de T1.
 if ((sum(sign(T1(:,1))) < 0) & (sum(sign(T1(:,2))) < 0))
   T1 = -T1;
 end
 s = sum(D1(1:p));
 for i = 1:p
   percent1(i) = (D1(i)/s)*100;
   acum1(i) = sum(percent1(1:i));
 end
% ----------------------------------------------------------
% Componentes principales a partir de la matriz de correlaciones.
% Ordenacion de los valores propios segun la variabilidad
% explicada ( de mas a menos). D2 es un vector fila.
% Las filas de T2 son los vectores propios ordenados.
%
 [T2,D2] = eigsort(R); T2 = T2';
% corregimos los signos de T2
 if ((sum(sign(T2(:,1))) < 0) & (sum(sign(T2(:,2))) < 0))
   T2 = -T2;
 end
 for i = 1:p
   percent2(i) = (D2(i)/p)*100;
   acum2(i) = sum(percent2(1:i));
 end
% ----------------------------------------------------------
% Las columnas de T1 son las componentes principales.
% Representacion de los datos.
 Y1 = X*T1;
 subplot(2,1,1);
 plot(Y1(:,1),Y1(:,2),'.b','MarkerSize',15)
 grid
 xlabel('1a. Componente Principal','FontSize',10)
 ylabel('2a. C.P.','FontSize',10)
 title(['A.C.P. a partir de S  (',num2str(acum1(2)),'%)'],...
       'FontSize',12)
 for i = 1:n,
   text(Y1(i,1),Y1(i,2),lab(i,:));
 end
% ----------------------------------------------------------
% Las columnas de T2 son las componentes principales
% (hay que estandarizar las variables).
 s = diag(sqrt(diag(S)));
% Representacion de los datos.
 Y2 = X*inv(s)*T2;
 subplot(2,1,2);
```

```
plot(Y2(:,1),Y2(:,2),'.b','MarkerSize',15)
grid
xlabel('1a Componente Principal','FontSize',10)
ylabel('2a. C.P.','FontSize',10)
title(['A.C.P. a partir de R  (',num2str(acum2(2)),'%)'],...
      'FontSize',12)
for i = 1:n,
  text(Y2(i,1),Y2(i,2),lab(i,:));
end
```

La función `eigsort.m` es una función auxiliar que se utiliza dentro de `comp.m` para ordenar las componentes principales según el porcentaje de variabilidad explicado (de mayor a menor).

```
% EIGSORT
%
% Funcion que ordena los valores propios segun el porcentaje
% de variabilidad explicada ( de mayor a menor ). Tambien se
% reordenan los vectores propios, segun los vap's.
% Nota: d es un vector columna.
%
function [v,d] = eigsort(a)
 [v,d] = eig(a);
 [x,i] = sort(-diag(real(d)));
 d = -x; v = v(:,i);
```

Sea `X` la matriz que contiene los datos de la Tabla 4.1. Para obtener las componentes principales haremos

```
[T1,Y1,acum1,T2,Y2,acum2] = comp(X)
```

La Figura 4.1 contiene la representación en componentes principales de estos países y el porcentaje de variabilidad explicado por las dos primeras componentes. Interpretaremos solamente las componentes calculadas a partir de $\mathbf{R}$, puesto que son las más adecuadas en este caso. Los coeficientes de estas dos componentes son las dos primeras columnas de la matriz `T2`. Los porcentajes de variabilidad acumulados se encuentran en el vector `acum`.

```
T2(:,1:2)=     0.3141    0.3924
              -0.3484    0.0414
               0.0735    0.1776
               0.4403    0.1340
               0.3297   -0.0834
              -0.1839   -0.0866
               0.1629    0.6398
              -0.0948   -0.3231
              -0.5218    0.2903
               0.3467   -0.3896
              -0.1006    0.1749

acum2(1:2)=   36.6353   54.1806
```

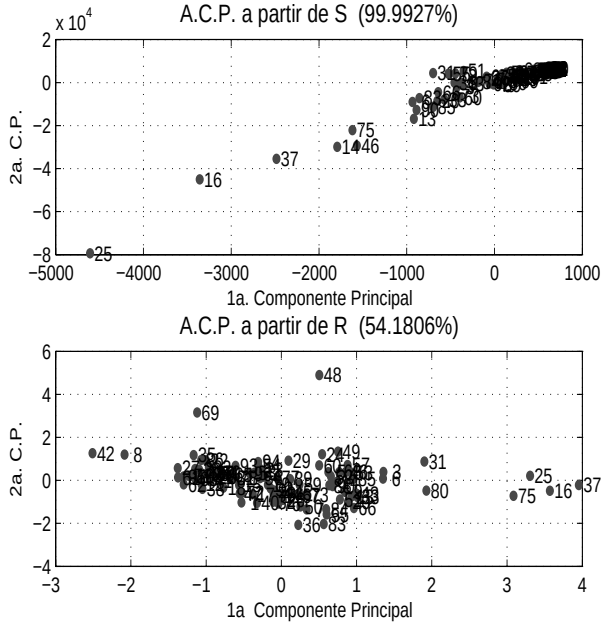


Figura 4.1.
Representación en componentes principales. (Problema 4.4.)

Las variables X_1, X_2, X_4, X_5, X_9 y X_{10} son las que más contribuyen en la primera componente principal, que puede interpretarse como un *índice de riqueza*. Mientras que X_1, X_7, X_8 y X_{10} son las que más contribuyen en la segunda componente, que podría interpretarse como un *índice de ruralidad*. Así por ejemplo, el grupo de países formados formado por Canadá (16), China (25), Francia (37) y Reino Unido (75) serían los más *ricos* según este índice que hemos construído, mientras que Bangladesh (8) y Haití (42) serían los más pobres. Por otro lado, Irán (48) y Pakistán (69) son los países con un índice de ruralidad más elevado, mientras que Finlandia (36) y Suecia (83) se encuentran en el lado opuesto.

PROBLEMA 4.5

En la Tabla 4.2 se recogen las siguientes variables medidas sobre 30 olmos hembra.

	nombre	*unidades*	*breve descripción*
X_1	*Longitud*	*mm*	*mayor medida de la corteza*
X_2	*Diámetro*	*mm*	*perpendicular a la longitud*
X_3	*Altura*	*mm*	*con madera dentro de la corteza*
X_4	*Peso total*	*g*	*todo el olmo*
X_5	*Peso desvainado*	*g*	*peso de la madera*
X_6	*Peso de las vísceras*	*g*	*peso de la tripa (después de sangrar)*
X_7	*Peso de la corteza*	*g*	*después de ser secado*

Este conjunto de datos pertenece a un estudio realizado por el Departamento de Industria Primaria y Pesca de Tasmania (Australia) en 1994. Los datos completos están disponibles en Nash et al. *(1994). Realícese un análisis de componentes principales e interprétense las dos primeras componentes.*

Tabla 4.2.
Datos para el Problema 4.5 (Fuente: Nash *et al.* 1994)

X1	X2	X3	X4	X5	X6	X7
0.53	0.42	0.135	0.677	0.2565	0.1415	0.21
0.53	0.415	0.15	0.7775	0.237	0.1415	0.33
0.545	0.425	0.125	0.768	0.294	0.1495	0.26
0.55	0.44	0.15	0.8945	0.3145	0.151	0.32
0.525	0.38	0.14	0.6065	0.194	0.1475	0.21
0.535	0.405	0.145	0.6845	0.2725	0.171	0.205
0.47	0.355	0.1	0.4755	0.1675	0.0805	0.185
0.44	0.34	0.1	0.451	0.188	0.087	0.13
0.565	0.44	0.155	0.9395	0.4275	0.214	0.27
0.55	0.415	0.135	0.7635	0.318	0.21	0.2
0.615	0.48	0.165	1.1615	0.513	0.301	0.305
0.56	0.44	0.14	0.9285	0.3825	0.188	0.3
0.58	0.45	0.185	0.9955	0.3945	0.272	0.285
0.68	0.56	0.165	1.639	0.6055	0.2805	0.46
0.68	0.55	0.175	1.798	0.815	0.3925	0.455
0.705	0.55	0.2	1.7095	0.633	0.4115	0.49
0.54	0.475	0.155	1.217	0.5305	0.3075	0.34
0.45	0.355	0.105	0.5225	0.237	0.1165	0.145
0.575	0.445	0.135	0.883	0.381	0.2035	0.26
0.45	0.335	0.105	0.425	0.1865	0.091	0.115
0.55	0.425	0.135	0.8515	0.362	0.196	0.27
0.46	0.375	0.12	0.4605	0.1775	0.11	0.15
0.525	0.425	0.16	0.8355	0.3545	0.2135	0.245
0.47	0.36	0.12	0.4775	0.2105	0.1055	0.15
0.5	0.4	0.14	0.6615	0.2565	0.1755	0.22
0.505	0.4	0.125	0.583	0.246	0.13	0.175
0.53	0.41	0.13	0.6965	0.302	0.1935	0.2
0.565	0.44	0.16	0.915	0.354	0.1935	0.32
0.595	0.495	0.185	1.285	0.416	0.224	0.485
0.475	0.39	0.12	0.5305	0.2135	0.1155	0.17

SOLUCIÓN

Sea `X` la matriz que contiene los datos de la Tabla 4.2. Mediante la instrucción

```
[T1,Y1,acum1,T2,Y2,acum2] = comp(X)
```

obtendremos las componentes principales. Aunque las unidades de medida de las variables son distintas, *mm* y *g*, las magnitudes son muy parecidas. Por tanto, nos quedaremos con las componentes calculadas a partir de la matriz de covarianzas, puesto que su interpretación es siempre más natural. Las dos primeras columnas de la matriz `T1` contienen estas dos componentes principales y el vector `acum1` contiene los porcentajes de variabilidad acumulados.

```
T1(:,1:2) =  0.1489    0.1339
            -0.0764   -0.0796
             0.4682    0.1373
             0.7550    0.2810
             0.1894    0.1829
             0.3825   -0.8629
             0.0221   -0.3162

acum1(1:2) = 97.6342   99.1793
```

La Figura 4.2 muestra la representación de los datos en función de las dos primeras componentes principales.

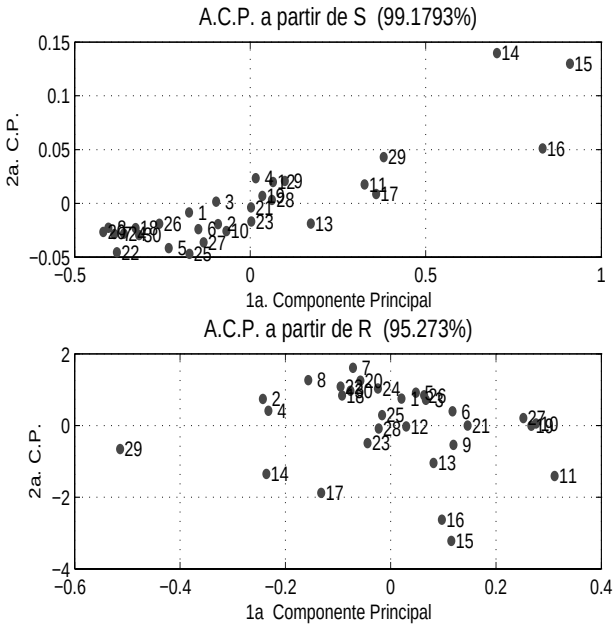


Figura 4.2.
Representación en componentes principales. (Problema 4.5.)

Puesto que el peso del diámetro, X_2, es muy pequeño en ambas componentes, no vamos a tener en cuenta esta variable a la hora de interpetar las componentes. La primera componente principal puede interpretarse como el tamaño del árbol, siendo el peso total, X_4, la variable que más contribuye en esta primera componente. La segunda componente principal puede interpretarse cómo la madera útil del árbol, puesto que el peso de las vísceras, X_6, y el peso de la corteza, X_7, tienen signo opuesto al resto de variables.
Observemos que, si hubiéramos calculado las componentes principales a partir de la matriz de correlaciones, la interpretación habría sido distinta.

PROBLEMA 4.6

En la Tabla 4.3 se recogen datos censales de algunos pueblos de España.

(a) *Determínense las dos primeras componentes principales a partir de la matriz de covarianzas* $\mathbf{S}$. *Calcúlese el porcentaje de varianza explicada por las dos primeras componentes.*

(b) *¿Qué se obtiene si se utiliza la matriz de correlaciones* $\mathbf{R}$ *en lugar de* $\mathbf{S}$*? Hállense los valores observados de las dos primeras componentes calculadas a partir de* $\mathbf{R}$.

Tabla 4.3.
Datos censales de pueblos de España (Problema 4.6)

Pueblo	Población total (en miles)	Inmigrantes extranjeros en últimos 5 años (en decenas)	Nº medio de hijos por pareja	Porcentaje de parados	Porcentaje de hogares con una sola persona
1	5.442	2.8	1.75	10.88	4.75
2	5.058	1.0	1.77	8.90	5.42
3	5.692	3.0	1.71	9.30	5.05
4	7.429	14.1	2.14	9.16	4.81
5	6.053	4.0	1.91	11.90	5.99
6	4.068	2.2	1.90	9.01	5.17
7	4.750	3.0	1.81	10.99	6.47
8	3.955	3.7	1.83	7.63	5.32
9	6.866	6.9	1.81	10.33	4.17
10	5.585	4.8	2.08	10.36	3.5
11	3.321	1.5	1.75	11.32	8.69
12	3.495	0.9	1.67	9.99	4.77
13	3.741	2.8	1.66	8.64	8.72
14	2.555	1.0	1.76	11.24	7.99

SOLUCIÓN

(a) Estamos observando en cada pueblo un vector $\mathbf{X} = (X_1, \ldots, X_5)'$, donde la variable X_1 es la población, la variable X_2 es el número de inmigrantes llegados en los últimos cinco años, etc. La matriz de covarianzas es

$$\mathbf{S} = \begin{pmatrix} 1.99 & 3.45 & 0.11 & 0.01 & -1.36 \\ & 11.68 & 0.33 & -0.59 & -1.93 \\ & & 0.02 & 0.00 & -0.10 \\ & & & 1.52 & 0.35 \\ & & & & 2.66 \end{pmatrix}.$$

Los autovalores de $\mathbf{S}$, así como el porcentaje de varianza total que explican las correspondientes componentes, se pueden ver a continuación:

Autovalor	Porcentaje $VT(\mathbf{S})$	Porcentaje acumulado
13.31	74.5	74.5
2.50	14.0	88.4
1.55	8.7	97.1
0.51	2.8	99.9
0.01	0.1	100

Las dos primeras componentes principales son:

$$\begin{aligned} Y_1 &= 0.31X_1 + 0.93X_2 + 0.03X_3 - 0.05X_4 - 0.21X_5 \\ Y_2 &= -0.30X_1 + 0.31X_2 - 0.01X_3 + 0.12X_4 + 0.89X_5. \end{aligned}$$

La primera componente es una media ponderada de la población y del número de inmigrantes recién llegados, así que hasta cierto punto mide la "vitalidad" (demográfica) de ese pueblo. La segunda componente está determinada en gran medida por el número de hogares compuestos por una sola persona.

(b) La matriz de correlaciones es

$$\mathbf{R} = \begin{pmatrix} 1 & 0.77 & 0.58 & 0.01 & -0.64 \\ & 1 & 0.74 & -0.15 & -0.37 \\ & & 1 & 0.00 & -0.47 \\ & & & 1 & 0.18 \\ & & & & 1 \end{pmatrix}.$$

Ya sabemos que las componentes principales de esta matriz no tienen por qué ser las mismas que las de $\mathbf{S}$. De hecho, los dos mayores autovalores de $\mathbf{R}$ son 2.82 y 1.03. Los correspondientes autovectores normalizados son

$$\begin{aligned} \mathbf{e}_1 &= (0.53, 0.52, 0.50, -0.08, -0.43)', \\ \mathbf{e}_2 &= (0.13, 0.03, 0.17, 0.96, 0.20)'. \end{aligned}$$

Como vemos, la primera componente calculada a partir de $\mathbf{R}$ representa en mayor medida las posibilidades de crecimiento de la población (lo que antes hemos llamado "vitalidad"). En cambio la segunda componente de $\mathbf{R}$ ahora está determinada por la proporción de parados.
Para calcular la matriz $\mathbf{W}$ de dimensión 14×2 con los valores observados de las dos primeras componentes, denotemos por $\mathbf{Z}$ la matriz 14×5 formada por las observaciones de $\mathbf{X}$ estandarizadas a media cero y varianza unidad. Además consideramos la matriz formada por los autovectores $\mathbf{e}_1$ y $\mathbf{e}_2$

$$\mathbf{A} = \begin{pmatrix} 0.53 & 0.52 & 0.50 & -0.08 & -0.43 \\ 0.13 & 0.03 & 0.17 & 0.96 & 0.20 \end{pmatrix}$$

Entonces $\mathbf{W} = \mathbf{Z}\,\mathbf{A}'$.

PROBLEMA 4.7

Sea $\mathbf{X}$ un vector aleatorio con matriz de correlaciones poblacionales

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}, \quad \textit{donde } 0 < \rho < \frac{1}{\sqrt{2}}.$$

(a) *Calcúlense los autovalores y autovectores de $\boldsymbol{\rho}$.*

(b) *Encuéntrense las componentes principales de $\boldsymbol{\rho}$.*

(c) *Calcúlese la proporción de variabilidad total explicada por las dos primeras componentes principales.*

(d) *Calcúlese la correlación entre la primera componente principal y X_2.*

SOLUCIÓN

(a) La ecuación característica de $\boldsymbol{\rho}$ tiene la expresión

$$|\boldsymbol{\rho} - \lambda \mathbf{I}| = \begin{vmatrix} 1-\lambda & \rho & 0 \\ \rho & 1-\lambda & \rho \\ 0 & \rho & 1-\lambda \end{vmatrix} = (1-\lambda)[(1-\lambda)^2 - 2\rho^2] = 0$$

Como ρ es positivo los autovalores ordenados de mayor a menor son $\lambda_1 = 1 + \sqrt{2}\rho$, $\lambda_2 = 1$ y $\lambda_3 = 1 - \sqrt{2}\rho$.
La ecuación característica, aplicada a λ_1, implica que

$$\begin{array}{rcl} -\sqrt{2}\rho x_1 + \rho x_2 & = & 0 \\ \rho x_1 - \sqrt{2}\rho x_2 + \rho x_3 & = & 0 \\ \rho x_2 - \sqrt{2}\rho x_3 & = & 0 \end{array} \Rightarrow \left\{ \begin{array}{l} x_1 = x_3 \\ x_2 = \sqrt{2}x_1 \end{array} \right.$$

Así que un autovector normalizado para λ_1 es $\mathbf{e}_1 = 1/2(1, \sqrt{2}, 1)'$. Análogamente calculamos los autovectores correspondientes a λ_2 y λ_3: $\mathbf{e}_2 = 1/\sqrt{2}(1, 0, -1)'$ y $\mathbf{e}_3 = 1/2(1, -\sqrt{2}, 1)'$.

(b) Dado que $E(\mathbf{X}) = \mathbf{0}$, las componentes principales centradas son

$$\begin{aligned} Y_1 &= \frac{1}{2}(X_1 + \sqrt{2}X_2 + X_3) \\ Y_2 &= \sqrt{2}(X_1 - X_3) \\ Y_3 &= \frac{1}{2}(X_1 - \sqrt{2}X_2 + X_3). \end{aligned}$$

(c) La variabilidad total es $VT(\boldsymbol{\rho}) = tr(\boldsymbol{\rho}) = 3$. La proporción de variabilidad total explicada por Y_1 es $(1 + \sqrt{2}\rho)/3$, luego cuánto más correladas están las variables, mejor resume Y_1 la información global. La proporción de variabilidad total explicada por Y_1 e Y_2 es $(2 + \sqrt{2}\rho)/3$.

(d) La correlación entre Y_1 y X_2 es $e_{12}\sqrt{\lambda_1}/\sqrt{\sigma_{22}} = \sqrt{(1 + \sqrt{2}\rho)/2}$.

PROBLEMA 4.8

Sea $\mathbf{X}$ *un vector aleatorio que sigue una distribución normal bivariante de media cero y matriz de covarianzas*

$$\boldsymbol{\Sigma} = \begin{pmatrix} 8 & 5 \\ 5 & 4 \end{pmatrix}.$$

(a) *Obténgase la función de densidad de* $\mathbf{X}$.

(b) *Realícese un análisis de componentes principales de* $\mathbf{X}$.

SOLUCIÓN

(a) La función de densidad de $\mathbf{X}$ es

$$f(\mathbf{x}) = \frac{1}{2\pi|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x}\right) = \frac{1}{2\sqrt{7}\pi} \exp\left(-\frac{1}{7}(2x_1^2 - 5x_1x_2 + 4x_2^2)\right),$$

y se encuentra representada en la Figura 4.3.

(b) Los autovalores de $\mathbf{\Sigma}$ son $\lambda_1 = 6 + \sqrt{29}$ y $\lambda_2 = 6 - \sqrt{29}$. Dos autovectores (no normalizados) de $\mathbf{\Sigma}$ correspondientes a estos autovalores son

$$\begin{aligned} \mathbf{v}_1 &= (5, \sqrt{29} - 2)' \simeq (5, 3.39)', \\ \mathbf{v}_2 &= (2 - \sqrt{29}, 5)' \simeq (-3.39, 5)', \end{aligned}$$

que normalizados dan $\mathbf{e}_1 = (0.83, 0.56)'$ y $\mathbf{e}_2 = (-0.56, 0.83)'$. Por tanto, las componentes principales de $\mathbf{X}$ son

$$\begin{aligned} Y_1 &= 0.83X_1 + 0.56X_2 \\ Y_2 &= -0.56X_1 + 0.83X_2. \end{aligned}$$

En la Figura 4.3 se observa que, en el caso de la distribución normal, las direcciones de las componentes principales coinciden con los ejes de las elipses que son los conjuntos de nivel de la densidad. Concretamente la dirección de la primera componente, es decir, la dirección sobre la que proyectaríamos $\mathbf{X}$ para que la proyección tuviera la máxima variabilidad, es precisamente el eje mayor de estas elipses.

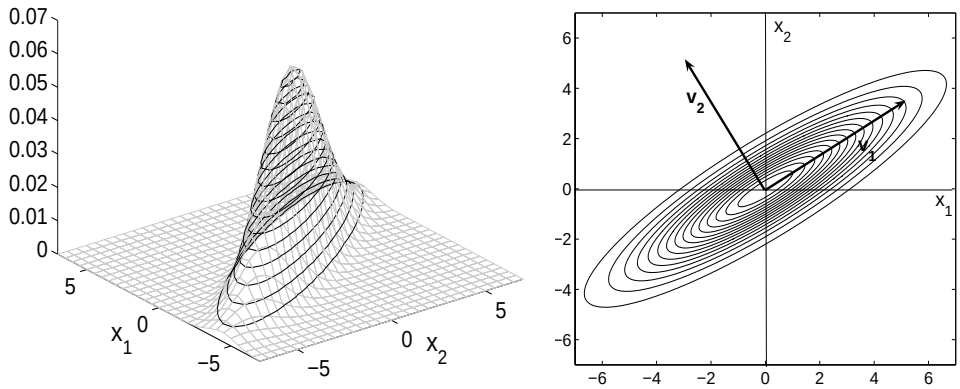


Figura 4.3.
Conjuntos de nivel de la densidad normal del Problema 4.8.

PROBLEMA 4.9

Sea

$$\mathbf{S} = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

la matriz de covarianzas muestral correspondiente al vector $\mathbf{X} = (X_1, X_2, X_3)'$*, donde* X_1 *representa la puntuación media en asignaturas de econometría para un alumno de la licenciatura conjunta en economía y derecho,* X_2 *es un promedio de sus resultados en asignaturas de derecho y* X_3 *es el resultado medio en asignaturas de libre elección.*

(a) *Calcúlense los autovalores de la matriz* $\mathbf{S}$.

(b) *Interprétese la segunda componente principal a partir de* $\mathbf{S}$ *sabiendo que su segundo autovector es*

$$\mathbf{e}_2 = (0.5744, -0.5744, 0.5744)'.$$

¿Cómo interpretaríamos el hecho de que un estudiante tenga un valor para la segunda componente principal mucho menor que el resto de sus compañeros?

(c) *¿Cuántas componentes principales son necesarias para explicar un mínimo de un* 80% *de la varianza? Escríbanse esas componentes en función de los datos originales* X_1*,* X_2 *y* X_3

SOLUCIÓN

(a) Los autovalores de $\mathbf{S}$ son $\lambda_1 \simeq 4.7$, $\lambda_2 = 3$ y $\lambda_3 \simeq 1.3$.

(b) La segunda componente principal enfrenta buenos resultados en econometría y asignaturas de libre elección con buenos resultados en derecho. Si la segunda componente para un estudiante es menor que las de sus compañeros quiere decir que se le dan mejor las asignaturas de derecho que aquellas de economía o las que escogiera en libre elección.

(c) Elegiremos dos componentes principales

$$\begin{aligned} Y_1 &= 0.5774X_1 + 0.7887X_2 + 0.2113X_3, \\ Y_2 &= 0.5774X_1 - 0.5774X_2 + 0.5774X_3, \end{aligned}$$

ya que la proporción de varianza acumulada explicada por la primera componente principal y por la primera y segunda componentes es 52.5% y 85.9% respectivamente.

PROBLEMA 4.10

Considérense dos variables aleatorias con media cero y matriz de covarianzas

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & 1 \\ 1 & \sigma^2 \end{pmatrix}$$

con $\sigma > 0$. *Se pide:*

(a) *Calcúlense los autovalores de* $\mathbf{\Sigma}$. *¿Para que valores de* σ *es* $\mathbf{\Sigma}$ *definida positiva?*

(b) *Encuéntrense las componentes principales a partir de* $\mathbf{\Sigma}$.

(c) *Calcúlese la proporción de variabilidad explicada por la primera componente.*

SOLUCIÓN

(a) La ecuación característica de $\mathbf{\Sigma}$ es

$$|\mathbf{\Sigma} - \lambda \mathbf{I}| = \lambda^2 - 2\sigma^2\lambda + \sigma^4 - 1 = 0\,.$$

Por tanto, los autovalores de $\mathbf{\Sigma}$ son $\lambda_1 = \sigma^2 + 1$ y $\lambda_2 = \sigma^2 - 1$. La matriz $\mathbf{\Sigma}$ es definida positiva cuando $\sigma > 1$.

(b) Los autovectores normalizados de $\mathbf{\Sigma}$ correspondientes a los autovalores λ_1 y λ_2 son, respectivamente, $\mathbf{e}_1 = \frac{1}{\sqrt{2}}(1,1)'$ y $\mathbf{e}_2 = \frac{1}{\sqrt{2}}(1,-1)'$. Entonces las componentes principales de $\mathbf{\Sigma}$ son

$$\begin{aligned} Y_1 &= \mathbf{e}_1'\mathbf{X} = \frac{1}{\sqrt{2}}(X_1 + X_2) \\ Y_2 &= \mathbf{e}_2'\mathbf{X} = \frac{1}{\sqrt{2}}(X_1 - X_2). \end{aligned}$$

Observemos que las componentes no dependen del parámetro σ. En cambio, su variabilidad sí depende de σ, como veremos a continuación.

(c) La varianza total es

$$VT(\mathbf{\Sigma}) = tr(\mathbf{\Sigma}) = 2\sigma^2\,.$$

La proporción de varianza total explicada por la primera componente es (véase también la Figura 4.4):

$$\frac{\lambda_1}{\mathrm{VT}(\mathbf{\Sigma})} = \frac{\sigma^2 + 1}{2\sigma^2} = \frac{1}{2} + \frac{1}{2\sigma^2}.$$

A medida que σ aumenta, la correlación entre X_1 y X_2 disminuye y una sola componente principal explica cada vez menos la variabilidad del sistema.

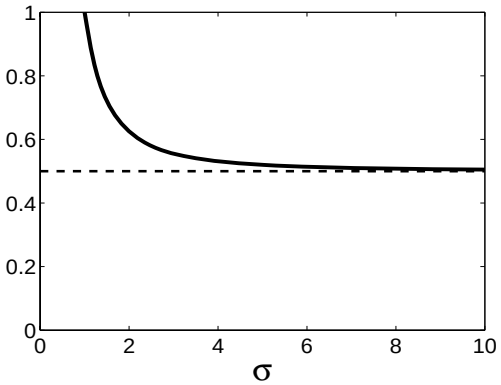


Figura 4.4.
Proporción de VT($\mathbf{\Sigma}$) explicada por Y_1 (Problema 4.10).

PROBLEMA 4.11

Considérese la matriz de varianzas-covarianzas de un vector aleatorio $\mathbf{X}$

$$\mathbf{\Sigma} = \begin{pmatrix} 9/2 & -3/2 \\ -3/2 & 9/2 \end{pmatrix}$$

(a) *Calcúlense las componentes principales de* $\mathbf{X}$ *a partir de* $\mathbf{\Sigma}$.

(b) *Considérese la siguiente matriz ortogonal*

$$\mathbf{A} = \begin{pmatrix} 2/\sqrt{5} & -1/\sqrt{5} \\ -1/\sqrt{5} & 2/\sqrt{5} \end{pmatrix}.$$

y calcúlense las componentes principales de $\mathbf{Y} = \mathbf{AX}$ *a partir de las componentes principales de* $\mathbf{X}$.

SOLUCIÓN

(a) Las componentes principales de $\mathbf{X}$ son $Z_1 = (X_1 - X_2)/\sqrt{2}$ y $Z_2 = (X_1 + X_2)/\sqrt{2}$.

(b) Dado que $\mathbf{A}$ es una matrix ortogonal, el vector $\mathbf{Y}$ no es más que una rotación rígida (o una reflexión) del vector $\mathbf{X}$. Como las direcciones de las componentes principales de $\mathbf{Y}$ son las direcciones de máxima variabilidad de $\mathbf{Y}$, para hallarlas lo único que tenemos que hacer es rotar las de $\mathbf{X}$. Otra manera de razonar es a partir de la descomposición espectral $\mathbf{\Sigma} = \mathbf{T}\,\mathbf{\Lambda}\,\mathbf{T}'$. Puesto que $\mathbf{Y} = \mathbf{A}\,\mathbf{X}$, entonces $var(\mathbf{Y}) = \mathbf{A}\,\mathbf{\Sigma}\,\mathbf{A}' = \mathbf{A}\,\mathbf{T}\,\mathbf{\Lambda}\,\mathbf{T}'\,\mathbf{A}'$. Es decir, para hallar los autovectores de $\mathbf{Y}$ hay que rotar los autovectores de $\mathbf{X}$ por $\mathbf{A}$, con lo que obtenemos $\mathbf{A}\mathbf{e}_1 = (1,-3)'/\sqrt{10}$ y $\mathbf{A}\mathbf{e}_2 = (3,1)'/\sqrt{10}$. Luego las componentes principales de $\mathbf{Y}$ son $W_1 = (X_1 - 3X_2)/\sqrt{10}$ y $W_2 = (3X_1 + X_2)/\sqrt{10}$.

PROBLEMA 4.12

Supongamos que dos observadores miden de manera independiente una variable aleatoria Z, pero cada uno de ellos comete un error de medida. Por esta razón las variables finalmente observadas son $X_1 = Z + \epsilon_1$ y $X_2 = Z + \epsilon_2$, donde ϵ_1 y ϵ_2 denotan los errores. Supongamos que $E(Z) = 7$, var$(Z) = 1$, $E(\epsilon_i) = 0$, var$(\epsilon_i) = \sigma^2$, $i = 1, 2$, donde σ es una constante positiva. Las variables Z, ϵ_1 y ϵ_2 son independientes entre sí. Denotemos $\mathbf{X} = (X_1, X_2)'$.

(a) *Calcúlense la esperanza y la matriz de varianzas-covarianzas de* $\mathbf{X}$.

(b) *Calcúlense las componentes principales de* $\mathbf{X}$.

(c) *Determínese, en función de σ, la proporción de variabilidad total explicada por las componentes principales. Interprétense los resultados obtenidos.*

SOLUCIÓN

(a) Dado que

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} Z + \epsilon_1 \\ Z + \epsilon_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} Z \\ \epsilon_1 \\ \epsilon_2 \end{pmatrix} = \mathbf{AY},$$

tenemos que

$$E(\mathbf{X}) = \mathbf{A}\ E(\mathbf{Y}) = (7, 7)'$$

y

$$Var(\mathbf{X}) = \mathbf{A}\ Var(\mathbf{Y})\ \mathbf{A}' = \begin{pmatrix} 1 + \sigma^2 & 1 \\ 1 & 1 + \sigma^2 \end{pmatrix}.$$

(b) Los autovalores de $Var(\mathbf{X})$ son $\lambda_1 = \sigma^2 + 2$ y $\lambda_2 = \sigma^2$. Las componentes principales son $Y_1 = (X_1 + X_2)/\sqrt{2}$ e $Y_2 = (X_1 - X_2)/\sqrt{2}$. Es decir, la primera componente promedia los datos proporcionados por ambos observadores.

(c) La proporción de VT explicada por Y_1 es

$$\frac{2 + \sigma^2}{2(1 + \sigma^2)} = \frac{1}{2} + \frac{1}{2(1 + \sigma^2)}.$$

Esto significa que, cuando la varianza de los errores aumenta, es decir, la incertidumbre en la observación de Z aumenta, se necesita en mayor medida la información proporcionada por ambos observadores. Mientras que, cuando σ^2 es baja, el promedio de las observaciones X_1 y X_2 es muy informativo acerca de Z.

PROBLEMA 4.13

La matriz de varianzas-covarianzas muestrales de unos datos bivariantes es

$$\mathbf{S} = \begin{pmatrix} 2647.5 & -530.9 \\ -530.9 & 127.4 \end{pmatrix}.$$

Los datos aparecen representados en el diagrama de dispersión de la Figura 4.5.

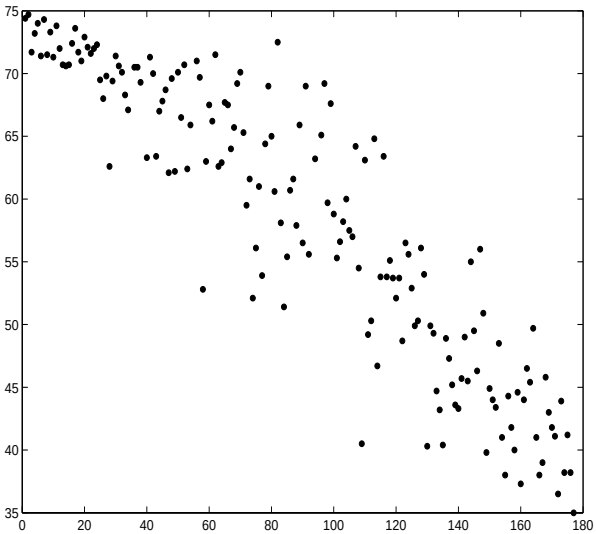


Figura 4.5.
Diagrama de dispersión (Problema 4.13)

(a) *Calcúlense las componentes principales a partir de* $\mathbf{S}$. *Interprétalas.*

(b) *Dibújese sobre el gráfico la dirección de la primera componente principal y explíquese qué significa intuitivamente esta dirección.*

(c) *Si cambio las unidades de medida de mis variables (por ejemplo, si las estandarizo) ¿varían las componentes principales con respecto a las de las variables originales?*

SOLUCIÓN

(a) Es sencillo comprobar que los autovalores de $\mathbf{S}$ son $\lambda_1 = 2754.8$ y $\lambda_2 = 20.1$ y que dos autovectores normalizados correspondientes a estos autovalores son respectivamente $\mathbf{e}_1 \simeq (-0.98, 0.2)'$ y $\mathbf{e}_2 \simeq (-0.2, -0.98)'$. Por tanto, las componentes principales de $\mathbf{S}$ son $Y_1 = -0.98X_1 - 0.2X_2$ e $Y_2 = -0.2X_1 - 0.98X_2$.

(b) La dirección de la primera componente principal Y_1 (o equivalentemente la dirección del autovector $\mathbf{e}_1$) es la dirección de mayor variabilidad en la muestra (ver Figura 4.6). Como Y_1 está determinada en gran medida por la variable X_1, la dirección de mayor variabilidad es aproximadamente la del eje de abscisas, aunque esto no se puede apreciar en la figura por las diferentes escalas de los ejes.

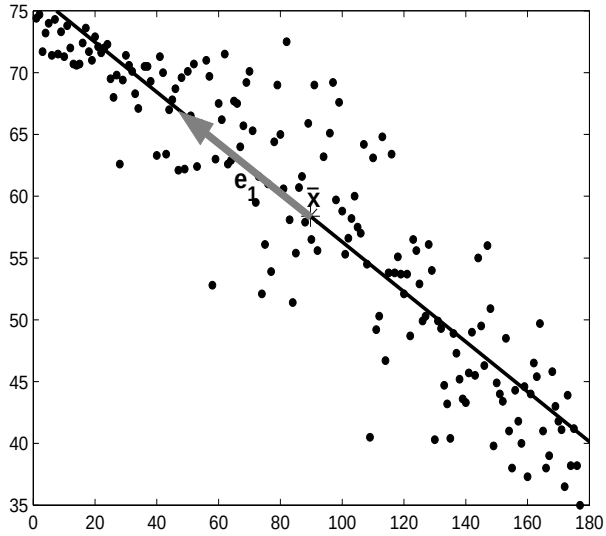


Figura 4.6.
Diagrama de dispersión y componentes principales (Problema 4.13)

Si dibujo una recta en la dirección de $\mathbf{e}_1$ y proyecto los datos sobre ella, las observaciones univariantes resultantes tendrán mayor varianza que las proyecciones en cualquier otra dirección. En este problema, como *Var*(X_1) es mucho mayor que *Var*(X_2), tenemos que Y_1 está dado principalmente por X_1. Es una situación clara en la que debemos estandarizar los datos, pues la diferencia entre variabilidades determina las componentes principales resultantes.

(c) Sí, ya sabemos que las componentes principales calculadas a partir de $\mathbf{R}$ no tienen por qué ser las mismas que las obtenidas a partir de $\mathbf{S}$.

PROBLEMA 4.14

Dados los pares de puntos (x_i, y_i), $1 \leq i \leq n$, de $\mathbb{R}^2$ demuéstrese que la recta de regresión que se obtiene por mínimos cuadrados ortogonales coincide con la primera componente principal.

SOLUCIÓN

Dado un conjunto de n puntos sobre $\mathbb{R}^2$, $\{(x_i, y_i), 1 \leq i \leq n\}$, las rectas de regresión con las que sin duda el lector estará más familiarizado son las que se obienen por mínimos cuadrados

ordinarios (recta de regresión lineal de Y sobre X y recta de regresión lineal de X sobre Y). En este ejercicio se trata de obtener una recta que sea una *buena aproximación* de la nube de puntos, pero *sin dar preferencia a ninguna coordenada*. A diferencia del caso de la regresión lineal, la función que debemos minimizar es la distancia, sobre la perpendicular, de los n pares puntos a una recta de ecuación $A\,x + B\,y = C$, donde (A, B) es su vector ortogonal. De hecho, para que el problema no sea indeterminado exigimos que este vector sea de norma unidad. De esta manera, la ecuación de la recta es $\alpha\,x + \beta\,y = \gamma$, donde $\alpha = A/\sqrt{A^2 + B^2}$, $\beta = B/\sqrt{A^2 + B^2}$, $\gamma = C/\sqrt{A^2 + B^2}$ (véase la Figura 4.7).

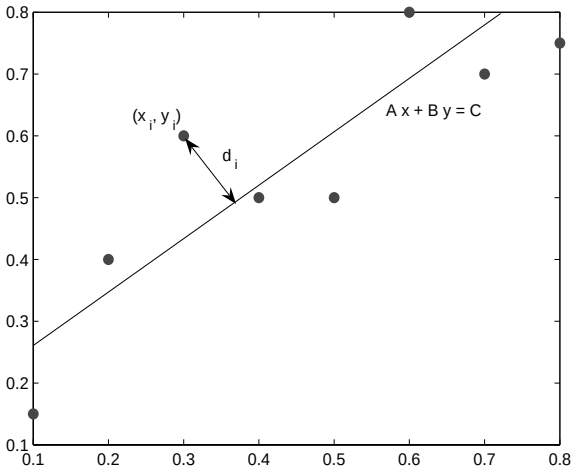


Figura 4.7.
Relación entre la regresión ortogonal y las componentes principales. (Problema 4.14.)

La función a minimizar es la media, $\overline{d^2}$, de las distancias ortogonales (al cuadrado), d_i^2, de los puntos (x_i, y_i) a la recta de ecuación $\alpha\,x + \beta\,y - \gamma = 0$,

$$\begin{aligned}\overline{d^2} &= \frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n}(\alpha\,x_i + \beta\,y_i - \gamma)^2 \\ &= \alpha^2\,\overline{x^2} + \beta^2\,\overline{y^2} + \gamma^2 + 2\alpha\beta\,\overline{xy} - 2\alpha\gamma\,\overline{x} - 2\beta\gamma\,\overline{y}.\end{aligned}$$

En función de las varianzas y covarianzas muestrales de (X, Y), hay que encontrar α, β y γ, tales que

$$\overline{d^2} = \alpha^2\,s_x^2 + \beta^2\,s_y^2 + 2\alpha\beta\,s_{xy} + (\alpha\,\overline{x} + \beta\,\overline{y} - \gamma)^2 \tag{4.1}$$

sea mínima. Puesto que la primera parte de (4.1) no depende de γ y $(\alpha\,\overline{x} + \beta\,\overline{y} - \gamma)^2 \geq 0$, se obtendrá el mínimo para $\gamma = \alpha\,\overline{x} + \beta\,\overline{y}$. Substituyendo este valor de γ en la ecuación de la recta,

$$\alpha(x - \overline{x}) + \beta(y - \overline{y}) = 0,$$

obtenemos que la recta de regresión pasa por el centro de gravedad de los puntos. Utilizando notación matricial, el problema de minimizar $\overline{d^2}$ es equivalente a encontrar los extremos de la

forma cuadrática

$$\overline{d^2} = \begin{pmatrix} \alpha & \beta \end{pmatrix} \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

El autovector (α, β) de autovalor máximo será el máximo de la función $\overline{d^2}$ y el autovector de autovalor mínimo será el mínimo que buscamos. Este último proporciona la dirección ortogonal a la recta $A\,x + B\,y = C$, mientras que el primero proporciona la dirección de la recta $A\,x + B\,y = C$. Así pues, esta recta es la dirección de máxima dispersión o variabilidad, o lo que es lo mismo, su vector director (que es el autovector de autovalor máximo) son los coeficientes de la primera componente principal. Asímismo la dirección ortogonal a la recta $A\,x + B\,y = C$ es la dirección de mínima dispersión, es decir, el vector ortogonal a esta recta (que es el autovector de autovalor mínimo) son los coeficientes de la segunda componente principal.

PROBLEMA 4.15

Determinar la edad de un árbol contando el número de anillos de una sección del tronco a través del microscopio es un trabajo muy laborioso. Por ello se busca la forma de predecir la edad de un árbol utilizando otras medidas más sencillas de obtener. La Tabla 4.4 contiene ocho variables medidas sobre 151 olmos. Las variables $X_1, \ldots, X_7$ son las mismas que las descritas en el Problema 4.5. La variable y es el número de anillos del olmo. Obténgase un modelo de regresión que permita predecir la edad de un olmo en función del resto de variables.

SOLUCIÓN

Sea `X` la matriz de datos que contiene las columnas de la Tabla 4.4 correspondientes a las variables $X_1, \ldots, X_7$ e `y` el vector columna que contiene la variable y. Queremos obtener un modelo de regresión lineal múltiple que permita predecir la edad del árbol (determinada por el número de anillos) en función de los regresores $X_1, \ldots, X_7$.
Si observamos la matriz `R` de correlaciones entre los regresores

```
R =
   1.0000   0.9889   0.9145   0.9234   0.9218   0.9017   0.8822
   0.9889   1.0000   0.9256   0.9285   0.9228   0.9058   0.8896
   0.9145   0.9256   1.0000   0.8996   0.8815   0.8979   0.8664
   0.9234   0.9285   0.8996   1.0000   0.9790   0.9350   0.9688
   0.9218   0.9228   0.8815   0.9790   1.0000   0.9455   0.9149
   0.9017   0.9058   0.8979   0.9350   0.9455   1.0000   0.8500
   0.8822   0.8896   0.8664   0.9688   0.9149   0.8500   1.0000
```

vemos que éstos están altamente correlacionados, por lo que es posible que existan problemas de multicolinealidad. Si calculamos el número de condición de la matriz `R`, es decir, la raíz cuadrada del cociente entre el máximo autovalor de `R` y el mínimo autovalor de `R`, vemos que es mayor que 30. Esto nos indica que `R` es una matriz mal condicionada y, efectivamente, vamos a tener problemas de multicolinealidad.

Tabla 4.4.
Datos del Problema 4.15

X_1	X_2	X_3	X_4	X_5	X_6	X_7	y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	y
0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	0.595	0.475	0.14	0.944	0.3625	0.189	0.315	9
0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	0.6	0.47	0.15	0.922	0.363	0.194	0.305	10
0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	0.555	0.425	0.14	0.788	0.282	0.1595	0.285	11
0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	0.615	0.475	0.17	1.1025	0.4695	0.2355	0.345	14
0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	0.575	0.445	0.14	0.941	0.3845	0.252	0.285	9
0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	0.62	0.51	0.175	1.615	0.5105	0.192	0.675	12
0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	0.52	0.425	0.165	0.9885	0.396	0.225	0.32	16
0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	0.595	0.475	0.16	1.3175	0.408	0.234	0.58	21
0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	0.58	0.45	0.14	1.013	0.38	0.216	0.36	14
0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	0.57	0.465	0.18	1.295	0.339	0.2225	0.44	12
0.525	0.38	0.14	0.6065	0.194	0.1475	0.21	14	0.625	0.465	0.14	1.195	0.4825	0.205	0.4 13	
0.43	0.35	0.11	0.406	0.1675	0.081	0.135	10	0.56	0.44	0.16	0.8645	0.3305	0.2075	0.26	10
0.49	0.38	0.135	0.5415	0.2175	0.095	0.19	11	0.46	0.355	0.13	0.517	0.2205	0.114	0.165	9
0.535	0.405	0.145	0.6845	0.2725	0.171	0.205	10	0.575	0.45	0.16	0.9775	0.3135	0.231	0.33	12
0.47	0.355	0.1	0.4755	0.1675	0.0805	0.185	10	0.565	0.425	0.135	0.8115	0.341	0.1675	0.255	15
0.5	0.4	0.13	0.6645	0.258	0.133	0.24	12	0.555	0.44	0.15	0.755	0.307	0.1525	0.26	12
0.355	0.28	0.085	0.2905	0.095	0.0395	0.115	7	0.595	0.465	0.175	1.115	0.4015	0.254	0.39	13
0.44	0.34	0.1	0.451	0.188	0.087	0.13	10	0.625	0.495	0.165	1.262	0.507	0.318	0.39	10
0.365	0.295	0.08	0.2555	0.097	0.043	0.1	7	0.695	0.56	0.19	1.494	0.588	0.3425	0.485	15
0.45	0.32	0.1	0.381	0.1705	0.075	0.115	9	0.665	0.535	0.195	1.606	0.5755	0.388	0.48	14
0.355	0.28	0.095	0.2455	0.0955	0.062	0.075	11	0.535	0.435	0.15	0.725	0.269	0.1385	0.25	9
0.38	0.275	0.1	0.2255	0.08	0.049	0.085	10	0.47	0.375	0.13	0.523	0.214	0.132	0.145	8
0.565	0.44	0.155	0.9395	0.4275	0.214	0.27	12	0.47	0.37	0.13	0.5225	0.201	0.133	0.165	7
0.55	0.415	0.135	0.7635	0.318	0.21	0.2	9	0.475	0.375	0.125	0.5785	0.2775	0.085	0.155	10
0.615	0.48	0.165	1.1615	0.513	0.301	0.305	10	0.36	0.265	0.095	0.2315	0.105	0.046	0.075	7
0.56	0.44	0.14	0.9285	0.3825	0.188	0.3	11	0.55	0.435	0.145	0.843	0.328	0.1915	0.255	15
0.58	0.45	0.185	0.9955	0.3945	0.272	0.285	11	0.53	0.435	0.16	0.883	0.316	0.164	0.335	15
0.59	0.445	0.14	0.931	0.356	0.234	0.28	12	0.53	0.415	0.14	0.724	0.3105	0.1675	0.205	10
0.605	0.475	0.18	0.9365	0.394	0.219	0.295	15	0.605	0.47	0.16	1.1735	0.4975	0.2405	0.345	12
0.575	0.425	0.14	0.8635	0.393	0.227	0.2	11	0.52	0.41	0.155	0.727	0.291	0.1835	0.235	12
0.58	0.47	0.165	0.9975	0.3935	0.242	0.33	10	0.545	0.43	0.165	0.802	0.2935	0.183	0.28	11
0.68	0.56	0.165	1.639	0.6055	0.2805	0.46	15	0.5	0.4	0.125	0.6675	0.261	0.1315	0.22	10
0.665	0.525	0.165	1.338	0.5515	0.3575	0.35	18	0.51	0.39	0.135	0.6335	0.231	0.179	0.2	9
0.68	0.55	0.175	1.798	0.815	0.3925	0.455	19	0.435	0.395	0.105	0.3635	0.136	0.098	0.13	9
0.705	0.55	0.2	1.7095	0.633	0.4115	0.49	13	0.495	0.395	0.125	0.5415	0.2375	0.1345	0.155	9
0.465	0.355	0.105	0.4795	0.227	0.124	0.125	8	0.465	0.36	0.105	0.431	0.172	0.107	0.175	9
0.54	0.475	0.155	1.217	0.5305	0.3075	0.34	16	0.435	0.32	0.08	0.3325	0.1485	0.0635	0.105	9
0.45	0.355	0.105	0.5225	0.237	0.1165	0.145	8	0.425	0.35	0.105	0.393	0.13	0.063	0.165	9
0.575	0.445	0.135	0.883	0.381	0.2035	0.26	11	0.545	0.41	0.125	0.6935	0.2975	0.146	0.21	11
0.355	0.29	0.09	0.3275	0.134	0.086	0.09	9	0.53	0.415	0.115	0.5915	0.233	0.1585	0.18	11
0.45	0.335	0.105	0.425	0.1865	0.091	0.115	9	0.49	0.375	0.135	0.6125	0.2555	0.102	0.22	11
0.55	0.425	0.135	0.8515	0.362	0.196	0.27	14	0.44	0.34	0.105	0.402	0.1305	0.0955	0.165	10
0.24	0.175	0.045	0.07	0.0315	0.0235	0.02	5	0.56	0.43	0.15	0.8825	0.3465	0.172	0.31	9
0.205	0.15	0.055	0.042	0.0255	0.015	0.012	5	0.405	0.305	0.085	0.2605	0.1145	0.0595	0.085	8
0.21	0.15	0.05	0.042	0.0175	0.0125	0.015	4	0.47	0.365	0.105	0.4205	0.163	0.1035	0.14	9
0.39	0.295	0.095	0.203	0.0875	0.045	0.075	7	0.385	0.295	0.085	0.2535	0.103	0.0575	0.085	7
0.47	0.37	0.12	0.5795	0.293	0.227	0.14	9	0.515	0.425	0.14	0.766	0.304	0.1725	0.255	14
0.46	0.375	0.12	0.4605	0.1775	0.11	0.15	7	0.37	0.265	0.075	0.214	0.09	0.051	0.07	6
0.325	0.245	0.07	0.161	0.0755	0.0255	0.045	6	0.36	0.28	0.08	0.1755	0.081	0.0505	0.07	6
0.525	0.425	0.16	0.8355	0.3545	0.2135	0.245	9	0.27	0.195	0.06	0.073	0.0285	0.0235	0.03	5
0.52	0.41	0.12	0.595	0.2385	0.111	0.19	8	0.375	0.275	0.09	0.238	0.1075	0.0545	0.07	6
0.4	0.32	0.095	0.303	0.1335	0.06	0.1	7	0.385	0.29	0.085	0.2505	0.112	0.061	0.08	8
0.485	0.36	0.13	0.5415	0.2595	0.096	0.16	10	0.7	0.535	0.16	1.7255	0.63	0.2635	0.54	19
0.47	0.36	0.12	0.4775	0.2105	0.1055	0.15	10	0.71	0.54	0.165	1.959	0.7665	0.261	0.78	18
0.405	0.31	0.1	0.385	0.173	0.0915	0.11	7	0.595	0.48	0.165	1.262	0.4835	0.283	0.41	17
0.5	0.4	0.14	0.6615	0.2565	0.1755	0.22	8	0.44	0.35	0.125	0.4035	0.175	0.063	0.129	9
0.445	0.35	0.12	0.4425	0.192	0.0955	0.135	8	0.325	0.26	0.09	0.1915	0.085	0.036	0.062	7
0.47	0.385	0.135	0.5895	0.2765	0.12	0.17	8	0.35	0.26	0.095	0.211	0.086	0.056	0.068	7
0.245	0.19	0.06	0.086	0.042	0.014	0.025	4	0.265	0.2	0.065	0.0975	0.04	0.0205	0.028	7
0.505	0.4	0.125	0.583	0.246	0.13	0.175	7	0.425	0.33	0.115	0.406	0.1635	0.081	0.1355	8
0.45	0.345	0.105	0.4115	0.18	0.1125	0.135	7	0.305	0.23	0.08	0.156	0.0675	0.0345	0.048	7
0.505	0.405	0.11	0.625	0.305	0.16	0.175	9	0.345	0.255	0.09	0.2005	0.094	0.0295	0.063	9
0.53	0.41	0.13	0.6965	0.302	0.1935	0.2	10	0.405	0.325	0.11	0.3555	0.151	0.063	0.117	9
0.425	0.325	0.095	0.3785	0.1705	0.08	0.1	7	0.375	0.285	0.095	0.253	0.096	0.0575	0.0925	9
0.52	0.4	0.12	0.58	0.234	0.1315	0.185	8	0.565	0.445	0.155	0.826	0.341	0.2055	0.2475	10
0.475	0.355	0.12	0.48	0.234	0.1015	0.135	8	0.55	0.45	0.145	0.741	0.295	0.1435	0.2665	10
0.565	0.44	0.16	0.915	0.354	0.1935	0.32	12	0.65	0.52	0.19	1.3445	0.519	0.306	0.4465	16
0.595	0.495	0.185	1.285	0.416	0.224	0.485	13	0.56	0.455	0.155	0.797	0.34	0.19	0.2425	11
0.475	0.39	0.12	0.5305	0.2135	0.1155	0.17	10	0.475	0.375	0.13	0.5175	0.2075	0.1165	0.17	10
0.31	0.235	0.07	0.151	0.063	0.0405	0.045	6	0.49	0.38	0.125	0.549	0.245	0.1075	0.174	10
0.555	0.425	0.13	0.7665	0.264	0.168	0.275	13	0.46	0.35	0.12	0.515	0.224	0.108	0.1565	10
0.4	0.32	0.11	0.353	0.1405	0.0985	0.1	8	0.28	0.205	0.08	0.127	0.052	0.039	0.042	9
0.595	0.475	0.17	1.247	0.48	0.225	0.425	20	0.175	0.13	0.055	0.0315	0.0105	0.0065	0.0125	5
0.57	0.48	0.175	1.185	0.474	0.261	0.38	11	0.17	0.13	0.095	0.03	0.013	0.008	0.01	4
0.605	0.45	0.195	1.098	0.481	0.2895	0.315	13	0.59	0.475	0.145	1.053	0.4415	0.262	0.325	15
0.6	0.475	0.15	1.0075	0.4425	0.221	0.28	15								

```
eig(R)= 0.0043   0.0105   0.0337   0.1080   0.1467   0.1972   6.4995

sqrt(max(eig(R))/min(eig(R)))=38.6608
```

Un procedimiento que se utiliza para el tratamiento de la multicolinealidad es transformar las variables mediante componentes principales, eliminar las menos informativas y expresar la variable respuesta en función de las componentes que resumen mayor variabilidad.
Para obtener las componentes principales utilizaremos la función `comp` y nos quedaremos con las dos primeras componentes, Y_1 e Y_2, calculadas a partir de la matriz de covarianzas, es decir, las dos primeras columnas de `Y1`, que explican el 98.7753% de la variabilidad de los datos. Dejamos para el lector la interpretación de estas dos primeras componentes.

```
[T1,Y1,acum1,T2,Y2,acum2]=comp(X)

T1 =
 0.2149   0.1781   0.0636   0.8445   0.3181   0.1705   0.2730
-0.6148  -0.4740  -0.1182   0.2667  -0.2090  -0.2970   0.4247
 0.4125   0.3411   0.0791  -0.1021  -0.4944  -0.3734   0.5594
 0.1264   0.0066  -0.2245  -0.0866   0.6127  -0.7412  -0.0346
-0.0792  -0.0492   0.1832  -0.4418   0.4805   0.3496   0.6398
-0.3997   0.3499   0.7945   0.0381   0.0641  -0.2508  -0.1347
-0.4730   0.7088  -0.5105  -0.0334   0.0142   0.0932   0.0575

acum1 =
97.5267  98.7753  99.5667  99.8136  99.9042  99.9606 100.0000
```

Para realizar la regresión lineal múltiple

$$y = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + u,$$

donde u es el término de perturbación aleatoria, construimos la matriz del diseño

```
regresores=[ones(151,1) Y1(1:2)]
```

y escribimos:

```
[b,bint,r,rint,stats] = regress(y,regresores)
rcoplot(r,rint)
```

que proporcionan el vector `b=(10.3907,-19.9826,11.7702)` de coeficientes β_j estimados. Por tanto, el modelo ajustado es $y = 10.3907 - 19.9826\,Y_1 + 11.7702\,Y_2$.
El vector `r` contiene los residuos del modelo y `rint` son los intervalos de confianza para los residuos. El vector `stats` contiene los resultados del contraste de significación del modelo, es decir, el valor del coeficiente de determinación R^2, el valor del estadístico F de Fisher y el p-valor asociado. La instrucción `rcoplot` permite obtener un gráfico de los residuos, junto con los intervalos de confianza al 95% (véase la Figura 4.8). Los triángulos son posibles *outliers*.

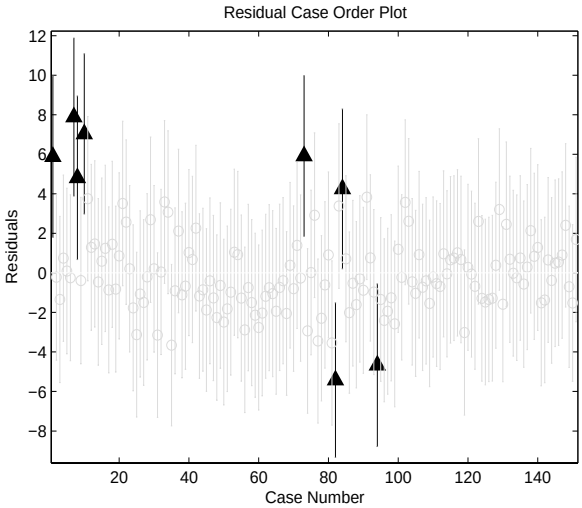


Figura 4.8.
Regresión en componentes principales. Gráfico de residuos (Problema 4.15.)

CAPÍTULO 5

Distancias estadísticas y escalado multidimensional (MDS)

La primera parte de este capítulo sirve de introducción para dos técnicas de representación de los individuos de un conjunto finito $\mathcal{E}$. Estas representaciones son de dos tipos:

- Una representación a lo largo de unos ejes reales que describe las analogías y diferencias entre los individuos (elementos de $\mathcal{E}$). Los ejes se interpretan como factores o causas de variabilidad y la información obtenida es de tipo *espacial*. Los problemas de la segunda parte de este Capítulo versan sobre este tema.

- Una representación como un grafo con estructura de árbol (dendrograma), como forma de representar clasificaciones jerárquicas entre los individuos. La información es de tipo *agrupativo*. Los problemas del Capítulo 6 tratan esta técnica.

El punto de partida en ambos casos es una matriz de distancias $\mathbf{D} = (\delta_{ij})$, de dimensión $n \times n$, siendo n el número de individuos del conjunto $\mathcal{E}$. Denotaremos por $\mathbf{D}^{(2)} = (\delta_{ij}^2)$, la matriz de cuadrados de distancias. El concepto de distancia entre objetos o individuos observados permite interpretar geométricamente muchas técnicas clásicas de Análisis Multivariante, equivalentes a representar estos objetos como puntos de un espacio métrico adecuado.

Similaridades, disimilaridades y distancias.
Una *disimilaridad* o *casi-métrica* es una función $\delta : \mathcal{E} \times \mathcal{E} \to \mathbb{R}_+$ tal que

- $\delta_{ij} = \delta_{ji}$, para todo i, j,

- $\delta_{ii} = 0$, para todo i.

Una *semi-métrica* es una disimilaridad que cumple

- $\delta_{ij} \leq \delta_{ik} + \delta_{kj}$, para todo i, j, k.

Una *métrica* es una semi-métrica que cumple

- $\delta_{ij} = 0 \Leftrightarrow i = j$, para todo i, j.

Una *ultramétrica* es una disimilaridad que cumple

- $\delta_{ij} \leq \max\{\delta_{ik}, \delta_{kj}\}$, para todo i, j, k.

En general, la palabra *distancia* puede hacer referencia tanto a una métrica como a una semi-métrica . Se usarán los términos técnicos de casi-métrica , semi-métrica y métrica cuando sea necesario precisar.
Una *similaridad* es una función $s : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ tal que

- $0 \leq s_{ij} \leq s_{ii} = 1$, para todo i, j,
- $s_{ij} = s_{ji}$, para todo i, j.

La siguiente transformación permite obtener una distancia de forma natural a partir de una similaridad s_{ij}:

$$\delta_{ij}^2 = s_{ii} + s_{jj} - 2\, s_{ij}. \tag{5.1}$$

Se dice que una distancia δ cumple la propiedad *euclídea* si existe una biyección $\psi : \mathcal{E} \to \mathbb{E} \subset \mathbb{R}^p$, para algún $p > 1$ tal que

$$\delta(x, y) = \|\psi(x) - \psi(y)\|, \quad \text{para todo } x, y \in \mathcal{E},$$

donde $\|\cdot\|$ es la norma euclídea en $\mathbb{R}^p$.

Escalado multidimensional métrico.
El escalado multidimensional métrico (o análisis de coordenadas principales) es una técnica de Análisis Multivariante cuyo objetivo es obtener una representación euclídea, exacta o aproximada, de los elementos de un conjunto $\mathcal{E}$ de n objetos, a partir de una matriz de disimilaridades $\mathbf{D}$ sobre $\mathcal{E}$.
Una *representación euclídea exacta* en dimensión $p \geq 1$ de $(\mathcal{E}, \mathbf{D})$ es un conjunto de n puntos $\mathbf{x}_1, \ldots, \mathbf{x}_n$ del espacio euclídeo $\mathbb{R}^p$, que verifica que las distancias euclídeas entre los $\mathbf{x}_i$ son iguales a los elementos correspondientes de la matriz $\mathbf{D}$.
En general una matriz de disimilaridades $\mathbf{D}$ no tiene ninguna representación euclídea exacta, a menos que $n = 2$. Cuando no es posible una representación exacta, o bien cuando la representación exacta es de dimensión grande, se hace necesario obtener una representación aproximada (de dimensión más reducida). Este aspecto se relaciona directamente con el problema de *reducción de la dimensión* estudiado en el Capítulo 4.

PROBLEMA 5.1

Se desea averiguar si una muestra de 20 individuos procede de una normal trivariante. Para ello se calculan las distancias de Mahalanobis de cada observación a la media muestral. En la Figura 5.1 se puede ver un qq-plot de estas distancias frente a cuantiles de la χ^2_3. ¿Qué se puede deducir del gráfico?

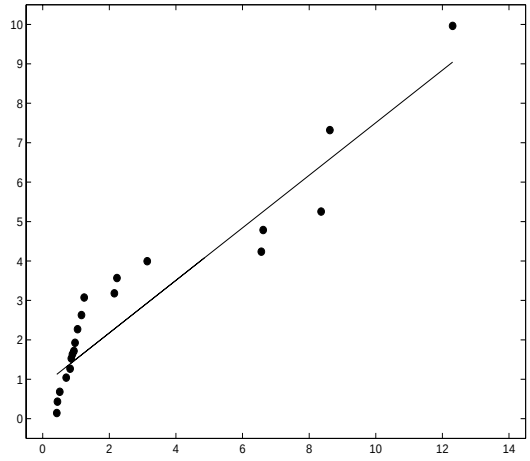


Figura 5.1.
qq-plot de distancias de Mahalanobis (Problema 5.1)

SOLUCIÓN

Si $\mathbf{X} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ entonces la distancia de Mahalanobis de $\mathbf{X}$ a su media verifica:

$$d^2_{Mah}(\mathbf{X}, \boldsymbol{\mu}) = (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2_3 \,.$$

Como $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son desconocidos los sustituimos por sus análogos muestrales, $\bar{\mathbf{x}}$ y $\mathbf{S}$, y tenemos que los cuadrados de las distancias de Mahalanobis de las observaciones $\mathbf{x}_i$ a la media $\bar{\mathbf{x}}$, dados por

$$(\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}),$$

deberían ser (aproximadamente) una muestra de una χ^2_3. Por tanto, los puntos del qq-plot se deberían ajustar a una línea recta. Como no es así, concluimos que la muestra no procede de una normal.

Observación. La recta de la Figura 5.1 no es la bisectriz del primer cuadrante. Esto es porque el gráfico fue generado con la orden `qqplot` de Matlab. Por ello la recta dibujada es la que une la pareja formada por los primeros cuartiles de ambas muestras con la pareja de terceros cuartiles.

PROBLEMA 5.2

La Tabla 4.1 contiene una serie de indicadores de distintos países del mundo. Calcúlese la matriz de distancias de Mahalanobis entre los 20 primeros países.

SOLUCIÓN

Construimos una función Matlab que calcule esta matriz de distancias.

```
% La funcion D=maha(X) calcula una matriz de cuadrados de
% distancias. El elemento (i,j) de la matriz D contiene el
% cuadrado de la distancia de Mahalanobis entre la fila "i"
% y la fila "j" de la matriz X.
%
% Entradas: una matriz X de dimension nxp.
% Salidas: una matriz D de dimension nxn.
%
function D = maha(X)
[n,p] = size(X);
% calculo del vector de medias y de la matriz de covarianzas
% de X:
S = cov(X,1);
% calculo de las distancias de Mahalanobis (al cuadrado):
D = zeros(n);
invS = inv(S);
for i = 1:n
   for j = i+1:n
      D(i,j) = (X(i,:)-X(j,:))*invS*(X(i,:)-X(j,:))';
   end
end
D = D+D';
```

Habíamos llamado `X` a la matriz de datos del Problema 4.4. Para obtener las distancias de Mahalanobis de los 20 primeros países haremos:

```
Z=X(1:20,:);
D=maha(Z);
```

Observando la matriz `D`, ¿qué países crees que son más parecidos? En el Problema 6.5 estudiaremos más detalladamente las semejanzas entre estos países.

PROBLEMA 5.3

Frecuentemente en las aplicaciones nos encontramos con una variable categórica nominal con k estados excluyentes medida sobre una muestra de $n = n_1 + \ldots + n_g$ individuos provenientes de g poblaciones. Se desea obtener una medida de disimilaridad entre estas poblaciones. En estas condiciones, el vector de frecuencias de cada población $\mathbf{n}_i = (n_{i1}, \ldots, n_{ik})$, para $i = 1, \ldots, g$, tiene una distribución conjunta multinomial con parámetros $(n_i, \mathbf{p}_i)$, donde $n_i = n_{i1} + \ldots + n_{ik}$ y $\mathbf{p}_i = (p_{i1}, \ldots, p_{ik})$. Dos medidas de disimilaridad son la distancia de Bhattacharyya, *conocida en genética como* distancia de Cavalli-Sforza, *cuya expresión es:*

$$d_{ij}^2 = \arccos\left(\sum_{l=1}^{k} \sqrt{p_{il}\, p_{jl}}\right)$$

y la distancia de Balakrishnan-Sanghvi *:*

$$d_{ij}^2 = 2\sum_{l=1}^{k} \frac{(p_{il} - p_{jl})^2}{p_{il} + p_{jl}}$$

La Tabla 5.1 contiene las proporciones génicas (observadas) de los grupos sanguíneos correspondientes a 10 poblaciones. Obténganse las distancias de Bhattacharyya y de Balakrishnan-Sanghvi entre estas poblaciones.

Tabla 5.1.
Proporciones génicas entre 10 poblaciones (Problema 5.3).

	Población	grupo A	grupo AB	grupo B	grupo O
1.	francesa	0.21	0.06	0.06	0.67
2.	checa	0.25	0.04	0.14	0.57
3.	germánica	0.22	0.06	0.08	0.64
4.	vasca	0.19	0.04	0.02	0.75
5.	china	0.18	0.00	0.15	0.67
6.	ainu	0.23	0.00	0.28	0.49
7.	esquimal	0.30	0.00	0.06	0.64
8.	negra USA	0.10	0.06	0.13	0.71
9.	española	0.27	0.04	0.06	0.63
10.	egipcia	0.21	0.05	0.20	0.54

SOLUCIÓN

Llamamos `X` a la Tabla 5.1, ya introducida en Matlab. Calculamos la matriz de cuadrados de distancias de Bhattacharyya de la siguiente forma:

```
q = sqrt(X);
DB2 = acos(q*q');
```

La matriz de cuadrados de distancias de Balakrishnan-Sanghvi se obtiene así:

```
[n,p] = size(X);
DBS2 = zeros(n);
for i = 1:n
   for j = 1:i-1
      if X(i,:)-X(j,:)==0,
         Y = 0;
      else
         Y = (X(i,:)-X(j,:))./sqrt(X(i,:)+X(j,:));
      end
     DBS2(i,j) = 2*Y*Y';
   end
end
DBS2 = DBS2+DBS2';
```

Por ejemplo, la matriz de cuadrados de distancias de Bhattacharyya es:

```
     0 0.1567 0.0435 0.1246 0.2863 0.3966 0.2622 0.1850 0.0800 0.2204
0.1567      0 0.1156 0.2665 0.2240 0.2605 0.2476 0.2093 0.1364 0.0897
0.0435 0.1156      0 0.1660 0.2715 0.3636 0.2608 0.1769 0.0778 0.1787
0.1246 0.2665 0.1660 0.0000 0.3221 0.4732 0.2607 0.2555 0.1517 0.3359
0.2863 0.2240 0.2715 0.3221      0 0.1933 0.1896 0.2710 0.2653 0.2491
0.3966 0.2605 0.3636 0.4732 0.1933      0 0.3101 0.3701 0.3642 0.2422
0.2622 0.2476 0.2608 0.2607 0.1896 0.3101      0 0.3608 0.2024 0.3226
0.1850 0.2093 0.1769 0.2555 0.2710 0.3701 0.3608 0.0000 0.2438 0.1997
0.0800 0.1364 0.0778 0.1517 0.2653 0.3642 0.2024 0.2438      0 0.2211
0.2204 0.0897 0.1787 0.3359 0.2491 0.2422 0.3226 0.1997 0.2211      0
```

Los individuos más cercanos (según la distancia de Battacharyya medida sobre sus proporciones génicas) son las poblaciones *francesa* y *germánica* con $\delta^2_{1,3} = 0.0435$, mientras que los más alejados son las poblaciones *francesa* y *ainu* con $\delta^2_{1,6} = 0.3966$. Estudiaremos con más detalle las proximidades entre estos individuos en los Problemas 5.9 y 6.3.

PROBLEMA 5.4

En muchas situaciones las variables que se observan sobre un conjunto de individuos son de naturaleza binaria. En estos casos para poder disponer de una matriz de distancias entre individuos se utilizan los coeficientes de similaridad.
El coeficiente de similaridad entre el individuo i y el individuo j, s_{ij}, se calcula a partir de las frecuencias:
$a =$ *"número de variables con respuesta* 1 *en ambos individuos",*
$b =$ *"número de variables con respuesta* 0 *en el primer individuo y con respuesta* 1 *en el segundo individuo",*
$c =$ *"número de variables con respuesta* 1 *en el primer individuo y con respuesta* 0 *en el segundo individuo",*
$d =$ *"número de variables con respuesta* 0 *en ambos individuos".*

Existen muchísimos coeficientes de similaridad (véase Cuadras 2004), pero los de Sokal-Michener y de Jaccard son especialmente interesantes porque dan lugar a una configuración euclídea (véase Problema 5.6). Se definen como:

$$\textit{Sokal y Michener:}\quad s_{ij}=\frac{a+d}{p},\quad \textit{Jaccard:}\quad s_{ij}=\frac{a}{a+b+c},$$

donde p es el número de variables observadas. Aplicando uno de estos coeficientes a un conjunto de n individuos se obtiene una matriz de similaridades $\mathcal{S}=(s_{ij})_{n\times n}$. Utilizando la fórmula (5.1) podemos obtener una distancia a partir de un coeficiente de similaridad . Este cálculo puede realizarse matricialmente:

$$\mathbf{D}^{(2)}=2\,(\mathbf{1}_n\,\mathbf{1}_n'-\mathcal{S}).$$

Se considera el siguiente conjunto de seis individuos formado por cinco animales, león, jirafa, vaca, oveja, gato doméstico, *junto con el* hombre. *Se miden seis variables binarias sobre estos individuos: X_1 =tiene cola, X_2 =es salvaje, X_3 =tiene el cuello largo, X_4 =es animal de granja, X_5 =es carnívoro, X_6 =camina sobre cuatro patas.*

(a) *Obténgase la matriz de datos.*

(b) *Calcúlense los coeficientes de similaridad de Sokal-Michener y de Jaccard para cada par de individuos y obténganse las matrices de distancias asociadas.*

SOLUCIÓN

(a) Consideremos el conjunto de individuos

$$\mathcal{E}=\{\textit{león, jirafa, vaca, oveja, gato doméstico, hombre}\},$$

entonces, la matriz de datos es

$$\begin{pmatrix}1&1&0&0&1&1\\1&1&1&0&0&1\\1&0&0&1&0&1\\1&0&0&1&0&1\\1&0&0&0&1&1\\0&0&0&0&1&0\end{pmatrix}\tag{5.2}$$

Observad que los individuos *vaca* y *oveja* puntúan igual, por lo que cualquier coeficiente de similaridad entre ellos deberá valer 1.

(b) Podemos construir dos funciones Matlab que calculen estos coeficientes de similaridad:

```
% SOKAL
%
% Dada una matriz de datos binarios X (n,p), la funcion S=sokal(X)
% devuelve la matriz de similaridades, segun el coeficiente de
% similaridad de Sokal y Michener, entre los n individuos.
%
function S = sokal(X)
 [n,p] = size(X);
 J = ones(n,p);
 a = X*X';
 d = (J-X)*(J-X)';
 S = (a+d)/p;

% JACCARD
%
% Dada una matriz de datos binarios X (n,p), la funcion
% S=jaccard(X) devuelve la matriz de similaridades, segun el
% coeficiente de similaridad de Jaccard, entre los n individuos.
%
function S = jaccard(X)
 [n,p] = size(X);
 J = ones(n,p);
 a = X*X';
 d = (J-X)*(J-X)';
 S = a./(p*ones(n)-d);
```

Si llamamos X a la matriz (5.2), las instrucciones en Matlab son:

```
[n,p] = size(X);
J = ones(n);
S_Sokal = sokal(X);
D2_Sokal = 2*(J-S_Sokal);
S_Jaccard = jaccard(X);
D2_Jaccard = 2*(J-S_Jaccard);
```

Por ejemplo, las matrices de similaridades son:

```
S_Sokal =

    1.0000    0.6667    0.5000    0.5000    0.8333    0.5000
    0.6667    1.0000    0.5000    0.5000    0.5000    0.1667
    0.5000    0.5000    1.0000    1.0000    0.6667    0.3333
    0.5000    0.5000    1.0000    1.0000    0.6667    0.3333
    0.8333    0.5000    0.6667    0.6667    1.0000    0.6667
    0.5000    0.1667    0.3333    0.3333    0.6667    1.0000

S_Jaccard =

    1.0000    0.6000    0.4000    0.4000    0.7500    0.2500
    0.6000    1.0000    0.4000    0.4000    0.4000         0
    0.4000    0.4000    1.0000    1.0000    0.5000         0
    0.4000    0.4000    1.0000    1.0000    0.5000         0
    0.7500    0.4000    0.5000    0.5000    1.0000    0.3333
    0.2500         0         0         0    0.3333    1.0000
```

Como ya se ha comentado anteriormente, el par de animales (*vaca*, *oveja*) es el más parecido con $s_{3,4}^{Sokal} = s_{3,4}^{Jaccard} = 1$. Les sigue el par (*león*, *gato*) con $s_{1,5}^{Sokal} = 0.8333$ y $s_{1,5}^{Jaccard} = 0.75$. En los Problemas 5.11 y 6.4 seguiremos estudiando las proximidades entre estos individuos.

PROBLEMA 5.5

Una situación muy habitual en análisis multivariante es disponer de un conjunto de datos mixto, es decir, un conjunto de individuos sobre los que se han observado tanto variables cuantitativas como cualitativas (o categóricas) . En estos casos es de gran utilidad la distancia de Gower, *cuyo cuadrado se define como* $d_{ij}^2 = 1 - s_{ij}$, *donde*

$$s_{ij} = \frac{\sum_{h=1}^{p_1} (1 - |x_{ih} - x_{jh}|/G_h) + a + \alpha}{p_1 + (p_2 - d) + p_3} \tag{5.3}$$

es el coeficiente de similaridad de Gower, p_1 *es el número de variables cuantitativas continuas,* p_2 *es el número de variables binarias,* p_3 *es el número de variables cualitativas (no binarias),* a *es el número de coincidencias* $(1, 1)$ *en las variables binarias,* d *es el número de coincidencias* $(0, 0)$ *en las variables binarias,* α *es el número de coincidencias en las variables cualitativas (no binarias) y* G_h *es el rango (o recorrido) de la* h*-ésima variable cuantitativa.*
Si $p_1 = p_3 = 0$ *entonces (5.3) coincide con el coeficiente de similaridad de Jaccard. Si se consideran las variables binarias como categóricas (es decir,* $p_1 = p_2 = 0$*) entonces (5.3) coincide con el coeficiente de similaridad de Sokal y Michener .*
La Tabla 5.2 contiene información sobre 50 jugadores de fútbol de la liga española (temporada 2006/07). Las variables observadas son:
X_1 *=número de goles marcados,* X_2 *=edad (años),* X_3 *=altura (m),* X_4 *=peso (kg),* X_5 *=pierna buena del jugador (*1 *=derecha,* 0 *=izquierda),* X_6 *=nacionalidad (*1 *=Argentina,* 2 *=Brasil,* 3 *=Camerún,* 4 *=Italia,* 5 *=España,* 6 *=Francia,* 7 *=Uruguay,* 8 *=Portugal,* 9 *=Inglaterra),* X_7 *=tipo de estudios (*1 *=sin estudios,* 2 *=básicos,* 3 *=medios,* 4 *=superiores).*
Obténgase la matriz de distancias de Gower entre estos individuos.

SOLUCIÓN

Una función Matlab que calcula el coeficiente de similaridad de Gower es:

```
% La funcion S=gower(X,p1,p2,p3,k) calcula una matriz de
%  similaridades, segun el coeficiente de similaridad de Gower.
%
% Entradas:
%   X   matriz de datos mixtos, cuyas columnas deben estar
%           ordenadas de la forma: continuas, binarias,
%           categoricas,
```

Tabla 5.2.
Variables observadas sobre jugadores de la liga española de fútbol 2006/07.

	Jugador	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1.	Ronaldinho	15	26	1.78	71	1	2	2
2.	Etoo	21	25	1.8	75	0	3	2
3.	Xavi	6	26	1.7	68	0	5	4
4.	Messi	7	19	1.69	67	0	1	3
5.	Puyol	1	28	1.78	78	0	5	3
6.	Raúl	7	29	1.8	73.5	1	5	3
7.	Ronaldo	18	30	1.83	82	0	2	1
8.	Beckham	4	31	1.8	67	0	9	3
9.	Casillas	0	25	1.85	70	0	5	4
10.	Cannavaro	0	33	1.76	75.5	0	4	2
11.	Torres	24	22	1.83	70	0	5	4
12.	Agüero	14	18	1.72	68	0	1	3
13.	Maxi	10	25	1.8	79	0	1	3
14.	Pablo	3	25	1.92	80	0	5	4
15.	Maniche	3	29	1.73	69	0	8	2
16.	Morientes	13	30	1.86	79	0	5	3
17.	Joaquín	5	25	1.79	75	0	5	4
18.	Villa	22	24	1.75	69	0	5	3
19.	Ayala	1	33	1.77	75.5	0	1	1
20.	Cañizares	0	36	1.81	78	1	5	3
21.	Jesús Navas	2	20	1.7	60	0	5	3
22.	Puerta	6	21	1.83	74	1	5	3
23.	Javi Navarro	7	32	1.82	75	0	5	3
24.	Daniel Alves	2	23	1.71	64	0	2	2
25.	Kanouté	12	29	1.92	82	1	6	1
26.	Valerón	9	31	1.84	71	0	5	3
27.	Arizmendi	8	22	1.92	78	0	5	3
28.	Capdevila	3	28	1.81	79	1	5	4
29.	Riki	7	26	1.86	80	0	5	3
30.	Coloccini	2	24	1.82	78	1	1	2
31.	Riquelme	10	28	1.82	75	0	1	2
32.	Forlán	17	27	1.72	75	0	7	3
33.	Cani	4	25	1.75	69.5	0	5	3
34.	Javi Venta	0	30	1.8	73	1	5	3
35.	Tachinardi	4	31	1.87	80	1	4	4
36.	Pandiani	6	30	1.84	74	0	7	1
37.	Tamudo	10	28	1.77	74	0	5	3
38.	De la Peña	2	30	1.69	69	0	5	3
39.	Luis García	8	25	1.8	68	0	5	3
40.	Jonathan	4	21	1.8	72	1	5	3
41.	Aimar	6	26	1.68	60	1	1	2
42.	Diego Milito	9	27	1.81	78	0	1	2
43.	Savio	3	32	1.71	68	1	2	2
44.	Sergio García	7	23	1.76	69	0	5	3
45.	Zapater	5	21	1.73	70.5	0	5	3
46.	Edu	6	27	1.82	74	1	2	3
47.	Juanito	2	30	1.83	80	0	5	4
48.	Melli	5	22	1.81	78	0	5	3
49.	Capi	7	29	1.75	73	0	5	2
50.	Doblas	0	25	1.84	78	0	5	3

```
%   p1  numero de variables continuas,
%   p2  numero de variables binarias,
%   p3  numero de variables categoricas (no binarias),
%   k   vector que contiene el numero de categorias de cada
%           variable categorica (no binaria) segun el orden
%           de entrada.
%
 function S = gower(X,p1,p2,p3,k)
 [n,p] = size(X);
% matriz de variables cuantitativas
 X1 = X(:,1:p1);
% matriz de variables binarias
 X2 = X(:,p1+1:p1+p2);
```

```
% matriz de variables categoricas
 X3 = X(:,p1+p2+1:p);
%
% calculos para las variables continuas
  rango = max(X1)-min(X1);
  for i = 1:n
     c(i,i) = p1;
     for j = 1:i-1
        c(i,j) = p1-sum(abs(X1(i,:)-X1(j,:))./rango);
        c(j,i) = c(i,j);
     end
  end
% calculo de las matrices a y d para las variables binarias
 J = ones(size(X2));
 a = X2*X2';
 d = (J-X2)*(J-X2)';
% calculos para las variables categoricas: cada variable
% categorica de k estados se transforma en k variables
% binarias que se yuxtaponen en una sola matriz Y1.
 Y1 = zeros(n,k(1));
 for i = 1:n
    Y1(i,X3(i,1)) = 1;
 end
 for j = 2:p3
    Y = zeros(n,k(j));
    for i = 1:n
       Y(i,X3(i,j)) = 1;
    end
    Y1 = [Y1 Y];
 end
 alpha = Y1*Y1';
 % calculo del coeficiente de similaridad de Gower
 S = (c+a+alpha)./(p*ones(n)-d);
```

Si llamamos `X` a la matriz que contiene los datos de la Tabla 5.2, las instrucciones para calcular las distancias de Gower son:

```
p1 = 4; p2 = 1; p3 = 2;
k = [9 4];
S_gower = gower(X,p1,p2,p3,k);
D2_gower = ones(size(S_gower))-S_gower;
```

Observando la matriz de cuadrados de distancias, ¿qué par de jugadores son más próximos? ¿qué par son más distantes? Estudiaremos con más detalle las proximidades entre estos jugadores en el Problema 5.10

PROBLEMA 5.6

Sea $\mathbf{D}$ *una matriz de distancias sobre* n *individuos de un conjunto* $\mathcal{E}$. *Se dice que* $(\mathcal{E}, \mathbf{D})$ *tiene (o admite) una* representación euclídea exacta *en dimensión* $p \geq 0$ *si existe un conjunto de* n *puntos* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *del espacio euclídeo* $\mathbb{R}^p$, *que verifica que las distancias euclídeas entre los* $\mathbf{x}_i$ $(i = 1, \ldots, n)$ *son iguales a los elementos correspondientes de la matriz* $\mathbf{D} = (\delta_{ij})_{1 \leq i,j \leq n}$, *es decir,*

$$\delta_{i,j}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j), \quad 1 \leq i, j \leq n.$$

Demuéstrese que $(\mathcal{E}, \mathbf{D})$ *tiene una representación euclídea de dimensión* $p \leq n - 1$ *si, y sólo si, la matriz*

$$\mathbf{B} = -\frac{1}{2}\,\mathbf{H}\,\mathbf{D}^{(2)}\,\mathbf{H} \tag{5.4}$$

es semidefinida positiva con $p = rg(\mathbf{B})$, *donde* $\mathbf{D}^{(2)}$ *denota la matriz de cuadrados de distancias y* $\mathbf{H}$ *es la matriz de centrado .*

SOLUCIÓN

$\Rightarrow$) Supongamos que $\mathbf{D}$ es euclídea, y sea

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

una configuración euclídea de $\mathbf{D}$ en $\mathbb{R}^p$. Los elementos de $\mathbf{D}$ (al cuadrado) son

$$\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) = -2\,a_{ij}. \tag{5.5}$$

Sea $\overline{\mathbf{x}}$ el centroide de $\mathbf{x}_1, \ldots, \mathbf{x}_n$, es decir,

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = \frac{1}{n}\mathbf{X}'\mathbf{1}_n. \tag{5.6}$$

Utilizaremos la siguiente notación:

$$\overline{a}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \quad \overline{a}_{\bullet j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}, \quad \overline{a}_{\bullet\bullet} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}.$$

Promediando (5.5) respecto de j

$$-2\,\overline{a}_{i\bullet} = \mathbf{x}_i'\mathbf{x}_i - 2\mathbf{x}_i'\overline{\mathbf{x}} + \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j'\mathbf{x}_j, \tag{5.7}$$

promediando (5.5) respecto de i

$$-2\,\overline{a}_{\bullet j} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i - 2\overline{\mathbf{x}}'\mathbf{x}_j + \mathbf{x}_j'\mathbf{x}_j, \tag{5.8}$$

y promediando la expresión (5.8) respecto de j

$$-2\,\overline{a}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i - 2\overline{\mathbf{x}}'\overline{\mathbf{x}} + \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j'\mathbf{x}_j. \tag{5.9}$$

Construimos la matriz

$$\mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{D}^{(2)}\mathbf{H} = \mathbf{H}\mathbf{A}\mathbf{H},$$

donde $\mathbf{A} = (a_{ij})_{1\le i,j\le n}$. Desarrollando esta expresión, se obtiene

$$\mathbf{B} = \mathbf{A} - \frac{1}{n}\mathbf{A}\mathbf{1}_n\mathbf{1}_n' - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\mathbf{A} + \frac{1}{n^2}\mathbf{1}_n\mathbf{1}_n'\mathbf{A}\mathbf{1}_n\mathbf{1}_n',$$

cuyos elementos son:

$$b_{ij} = a_{ij} - \overline{a}_{i\bullet} - \overline{a}_{\bullet j} + \overline{a}_{\bullet\bullet}. \tag{5.10}$$

Substituyendo las expresiones (5.7), (5.8) y (5.9) en (5.10) y operando, se obtiene

$$b_{ij} = \mathbf{x}_i'\mathbf{x}_j - \mathbf{x}_i'\overline{\mathbf{x}} - \overline{\mathbf{x}}'\mathbf{x}_j + \overline{\mathbf{x}}'\overline{\mathbf{x}} = (\mathbf{x}_i - \overline{\mathbf{x}})'(\mathbf{x}_j - \overline{\mathbf{x}}).$$

Matricialmente, utilizando la expresión (5.6), tenemos que:

$$\mathbf{B} = (\mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}')(\mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}')' = (\mathbf{H}\mathbf{X})(\mathbf{H}\mathbf{X})',$$

de donde se deduce que $\mathbf{B} \ge 0$ y $rg(\mathbf{B}) = p$, puesto que $rg(\mathbf{H}\mathbf{X}) = p$.

$\Leftarrow$) Supongamos que $\mathbf{B} \ge 0$ con $p = rg(\mathbf{B})$. Entonces, según el teorema de descomposición espectral,

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}',$$

donde $\mathbf{U}$ es una matriz ortogonal y $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$.
Definiendo

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2},$$

se tiene que

$$\mathbf{B} = \mathbf{X}\mathbf{X}',$$

cuyos elementos son $b_{ij} = \mathbf{x}_i'\mathbf{x}_j$.
Veamos que los puntos representados por $\mathbf{x}_1, \ldots, \mathbf{x}_n$ son una configuración euclídea de $\mathbf{D}$. Utilizando las expresiones (5.5) y (5.10),

$$\begin{aligned}(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) &= \mathbf{x}_i'\mathbf{x}_i - 2\,\mathbf{x}_i'\mathbf{x}_j + \mathbf{x}_j'\mathbf{x}_j = b_{ii} - 2\,b_{ij} + b_{jj}\\ &= a_{ii} - 2\,a_{ij} + a_{jj} = -2\,a_{ij} = \delta_{ij}^2,\end{aligned}$$

puesto que $a_{ii} = -\delta_{ii}^2/2 = 0$, $a_{jj} = -\delta_{jj}^2/2 = 0$.
El rango de $\mathbf{B}$ es siempre menor o igual que $n-1$, puesto que $\mathbf{1}_n$ es un autovector de $\mathbf{B}$ cuyo autovalor es 0, es decir, $\mathbf{B}\mathbf{1}_n = \mathbf{H}\mathbf{A}\mathbf{H}\mathbf{1}_n = \mathbf{0}$.

PROBLEMA 5.7

Demuéstrese que, si la matriz $\mathbf{B}$ *definida en (5.4) tiene autovalores negativos, la transformación*

$$\tilde{\delta}_{ij}^2 = \begin{cases} \delta_{ij}^2 + c, & i \neq j, \\ 0, & i = j, \end{cases} \tag{5.11}$$

donde $c \geq 2|\lambda|$*, siendo* λ *el autovalor negativo de módulo máximo, da lugar a una nueva matriz de distancias* $\tilde{\mathbf{D}}$ *que admite una representación euclídea. Esta transformación se denomina* q-aditiva *y es la que menos distorsiona la distancia original. Los programas de escalado multidimensional (en inglés,* MDS *o* multidimensional scaling*) utilizan otras transformaciones no lineales más complicadas (véanse Mardia, Kent y Bibby 1979, Peña 2002, Cuadras 2004).*

SOLUCIÓN

Sean $\lambda_1 > \ldots > \lambda_k > 0 > \lambda'_1 > \ldots > \lambda'_m$, con $m + k = n - 1$ los autovalores de la matriz $\mathbf{B}$. Sea $\mathbf{D}^{(2)} = (\delta_{ij}^2)$ la matriz de cuadrados de distancias y $\tilde{\mathbf{D}}^{(2)} = (\tilde{\delta}_{ij}^2)$ la matriz de cuadrados de distancias transformada según (5.11), que matricialmente se escribe

$$\tilde{\mathbf{D}}^{(2)} = \mathbf{D}^{(2)} + c\,(\mathbf{1}_n\mathbf{1}'_n - \mathbf{I}).$$

La matriz $\tilde{\mathbf{B}}$ correspondiente es

$$\tilde{\mathbf{B}} = -\frac{1}{2}\mathbf{H}\tilde{\mathbf{D}}^{(2)}\mathbf{H} = -\frac{1}{2}\mathbf{H}\mathbf{D}^{(2)}\mathbf{H} - \frac{c}{2}\mathbf{H}(\mathbf{1}_n\mathbf{1}'_n - \mathbf{I})\mathbf{H} = \mathbf{B} + \frac{c}{2}\mathbf{H},$$

puesto que $\mathbf{H}\mathbf{1}_n = \mathbf{0}$ y $\mathbf{H}^2 = \mathbf{H}$. Si $\mathbf{v}$ es un autovector de la matriz $\mathbf{B}$ de autovalor no nulo λ, es decir, $\mathbf{B}\,\mathbf{v} = \lambda\,\mathbf{v}$, entonces:

$$\begin{aligned} \tilde{\mathbf{B}}\,\mathbf{v} &= (\mathbf{B} + \frac{c}{2}\mathbf{H})\,\mathbf{v} = \mathbf{B}\,\mathbf{v} + \frac{c}{2}\mathbf{H}\,\mathbf{v} \\ &= \lambda\,\mathbf{v} + \frac{c}{2}(\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}'_n)\,\mathbf{v} = \lambda\,\mathbf{v} + \frac{c}{2}\,\mathbf{v} = \left(\lambda + \frac{c}{2}\right)\,\mathbf{v}, \end{aligned}$$

puesto que $\mathbf{1}'_n\,\mathbf{v} = 0$, al ser $\mathbf{1}_n$ autovector de $\mathbf{B}$ de autovalor 0. Por tanto, si λ'_m es el autovalor de $\mathbf{B}$ negativo de módulo máximo, entonces

$$\lambda'_m + \frac{c}{2} \geq 0 \Leftrightarrow c \geq -2\lambda'_m = 2|\lambda'_m|.$$

En particular, si $c = 2|\lambda'_m|$ la transformación es euclídea en dimensión $m + k - 1$, puesto que existen $m + k - 1$ autovalores positivos y un autovalor nulo.
Es interesante disponer de una función Matlab que realice esta transformación.

```
% non2euclid
%
% Dada una matriz D (nxn) de cuadrados de distancias
% no euclidea, la funcion D1=non2euclid(D) devuelve
% una matriz D1 de cuadrados de distancias euclidea.
```

```
%
function D1 = non2euclid(D)
[n,n] = size(D);
H = eye(n)-ones(n)/n;
[T,Lambda] = eig(-H*D*H/2);
m = min(diag(Lambda));
D1 = D-2*m*ones(n)+2*m*eye(n);
```

PROBLEMA 5.8

Sea $\mathcal{E}$ un conjunto de n individuos cuya matriz euclídea de distancias es $\mathbf{D}$ y cuya representación en coordenadas principales es $\mathbf{X}$. Se desean obtener las coordenadas de un nuevo individuo, al que llamaremos individuo $n+1$, del cual se conocen los cuadrados de sus distancias a los n individuos del conjunto $\mathcal{E}$. Si $\mathbf{d} = (\delta^2_{n+1,1}, \ldots, \delta^2_{n+1,n})'$ es el vector columna que contiene las distancias al cuadrado del individuo $n+1$ a los restantes, demuéstrese que la representación en coordenadas principales del individuo $n+1$ viene dada por

$$\mathbf{x}_{n+1} = \frac{1}{2}\mathbf{\Lambda}^{-1}\,\mathbf{X}'\,(\mathbf{b}-\mathbf{d}), \tag{5.12}$$

donde $\mathbf{b} = diag(\mathbf{B}) = (b_{11}, \ldots, b_{nn})'$, $\mathbf{B} = \mathbf{X}\,\mathbf{X}' = \mathbf{U}\,\mathbf{\Lambda}\,\mathbf{U}'$ y $\mathbf{U}$ es una matriz ortogonal. La ecuación (5.12) se conoce como fórmula de interpolación de Gower *(Gower 1968).*

SOLUCIÓN

La distancia (al cuadrado) del individuo $n+1$ a un individuo i de $\mathcal{E}$ es:

$$\delta^2_{n+1,i} = (\mathbf{x}_{n+1}-\mathbf{x}_i)'(\mathbf{x}_{n+1}-\mathbf{x}_i) = \mathbf{x}'_{n+1}\mathbf{x}_{n+1} - 2\mathbf{x}'_{n+1}\mathbf{x}_i + \mathbf{x}'_i\mathbf{x}_i,$$

para $1 \leq i \leq n$. Matricialmente,

$$\mathbf{d} = \|\mathbf{x}_{n+1}\|^2\,\mathbf{1}_n - 2\,\mathbf{X}\,\mathbf{x}_{n+1} + \mathbf{b}.$$

Operando y multiplicando por la izquierda por $\mathbf{X}'$, tenemos que:

$$2\,\mathbf{X}'\,\mathbf{X}\,\mathbf{x}_{n+1} = \mathbf{X}'\,(\mathbf{b}-\mathbf{d}) + \|\mathbf{x}_{n+1}\|^2\,\mathbf{X}'\mathbf{1}_n.$$

y puesto que $\mathbf{X}'\,\mathbf{1}_n = \mathbf{0}$ y $\mathbf{X} = \mathbf{U}\,\mathbf{\Lambda}^{1/2}$,

$$\mathbf{x}_{n+1} = \frac{1}{2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{b}-\mathbf{d}) = \frac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{X}'(\mathbf{b}-\mathbf{d}).$$

PROBLEMA 5.9

Obténgase una representación en coordenadas principales de las poblaciones del Problema 5.3, utilizando la matriz de distancias de Bhattacharyya. ¿Cuál es la dimensión de la representación euclídea? Determínese cuál es el porcentaje de variabilidad explicado por las dos primeras coordenadas principales.

SOLUCIÓN

Construimos una función Matlab para obtener las coordenadas principales a partir de una matriz de cuadrados de distancias.

```
% COORP
%
%
% La funcion [X,vaps,percent,acum] = coorp(D) calcula las
% coordenadas principales a partir de una matriz de D de cuadrados
% distancias,
%
%
% Entradas:  D = matriz de cuadrados de distancias.
%
%
% Devuelve:
%    X = matriz de coordenadas principales,
%    vaps = vector fila que contiene los autovalores,
%    percent = vector fila que contiene los porcentajes de
%              variabilidad explicados por cada coordenada.
%    acum = vector fila que contiene los porcentajes de
%              variabilidad acumulados.
%
%
 function [X,vaps,percent,acum] = coorp(D)
 [n,n] = size(D);
% comprobamos que D es euclidea (ie, B>=0)
 H = eye(n)-ones(n)/n;
 B = -H*D*H/2;
 L = eig(B);
 m = min(L);
 epsilon = 1.e-6;
 if abs(m) < epsilon
    % hacemos la transformacion non2euclid
    D1 = non2euclid(D);
    B = -H*D1*H/2;
 end
%--------------------------------------------------
% calculo de las coordenadas principales (solo consideramos
% las no nulas)
%
 [T,Lambda,V] = svd(B);
```

```
 vaps = diag(Lambda)';
 j = 1;
 while vaps(j)>epsilon
    T1 = T(:,1:j);
    X = T1*sqrt(Lambda(1:j,1:j));
    j = min(j+1,n);
 end
 percent = vaps/sum(vaps)*100;
 acum = zeros(1,n);
 for i = 1:n
    acum(i) = sum(percent(1:i));
 end
%-------------------------------------------------------
% vector de etiquetas para los individuos
%
 for i = 1:n
   lab(i,:) = sprintf('%3g',i);
 end
%-------------------------------------------------------
% representacion de los datos en dimension 2
 plot(X(:,1),X(:,2),'.b','MarkerSize',15)
 grid
 xlabel('Primera coordenada principal','FontSize',10)
 ylabel('Segunda coordenada principal','FontSize',10)
 title(['Porcentaje de variabilidad explicada ', ...
                num2str(acum(2)),'%'],'FontSize',12)
 for i = 1:n,
    text(X(i,1),X(i,2),lab(i,:));
 end
```

En el Problema 5.3 habíamos calculado la matriz `DB2` de cuadrados de distancias de Bhattacharyya entre las poblaciones. Para obtener la representación en coordenadas principales haremos:

```
[X,vaps,percent,acum] = coorp(DB2);
```

La dimensión de la representación euclídea es el número de autovalores no nulos, es decir, la dimensión del vector fila `vaps`. El vector fila `percent` contiene los porcentajes de variabilidad explicados por cada coordenada principal y el vector `acum` contiene los porcentajes acumulados. En la Figura 5.2 se encuentra la representación de las poblaciones {*francesa, checa, germánica, vasca, china, ainu, esquimal, negra USA, española, egipcia*} en función de las dos primeras coordenadas principales. A grandes rasgos pueden distinguirse tres grupos, que estudiaremos con más detalle en el Problema 6.3.

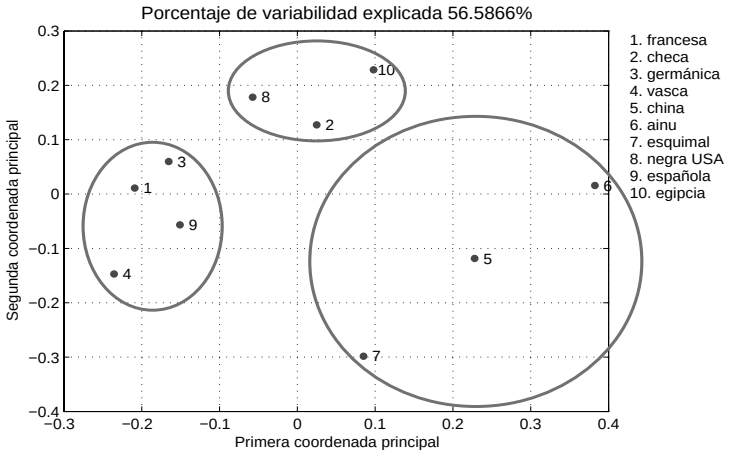


Figura 5.2.
Representación en coordenadas principales (Problema 5.9).

PROBLEMA 5.10

Utilizando la matriz de distancias del Problema 5.5 obténgase una representación de los jugadores en coordenadas principales. Determínese cuál es el porcentaje de variabilidad explicado por las dos primeras coordenadas principales . ¿Qué se puede decir de las semejanzas entre jugadores?

SOLUCIÓN

En el Problema 5.5 habíamos obtenido la matriz de cuadrados de distancias `D2_gower`. Utilizando la función `coorp` construida en el Problema 5.9 realizaremos la representación en coordenadas principales:

```
[Y,vaps,percent,acum] = coorp(D2_gower);
```

La Figura 5.3 contiene la representación de los jugadores en función de las dos primeras coordenadas principales. Quizá al lector le resulte entretenido buscar parecidos entre distintos jugadores.

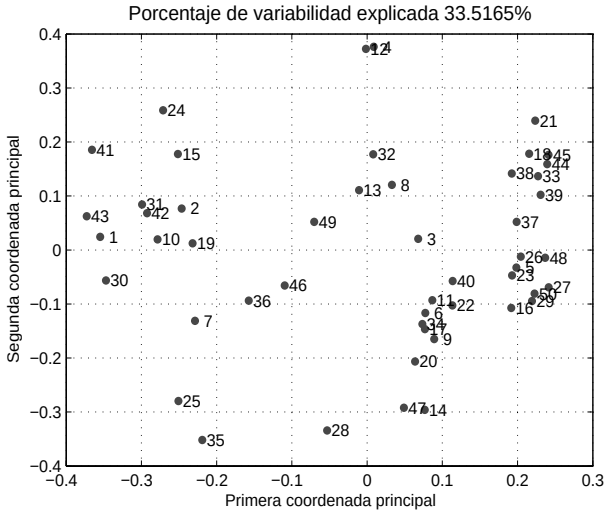


Figura 5.3.
Representación en coordenadas principales (Problema 5.10).

PROBLEMA 5.11

Para los datos del Problema 5.4

(a) *obténgase una representación en coordenadas principales utilizando la matriz de distancias calculada a partir del coeficiente de similaridad de Sokal y Michener.*

(b) *Sin volver a recalcular las coordenadas principales, añádase el* elefante *al conjunto de animales y obténganse sus coordenadas (véase el Problema 5.8).*

SOLUCIÓN

(a) Habíamos llamado `X` a la matriz de datos del Problema 5.4. Llamaremos `Y` a la matriz que contiene las coordenadas principales del conjunto de animales.

```
[Y,vaps,percent,acum] = coorp(D2_Sokal);
```

La Figura 5.4 contiene la reprentación de los animales en función de las dos primeras coordenadas principales. Podemos observar que el primer eje ordena a los animales dependiendo de si son carnívoros o herbívoros, mientras que el segundo eje ordena a los animales en función de que sean salvajes o no. En el Problema 6.4 estudiaremos con más detalle las agrupaciones entre estos individuos.

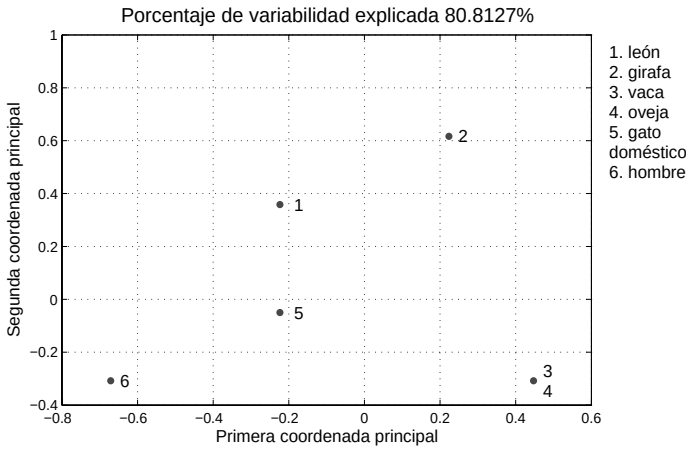


Figura 5.4.
Representación en coordenadas principales (Problema 5.11).

(b) Las puntuaciones del *elefante* según las variables del Problema 5.4 son (1 1 0 0 0 1). Recordemos que habíamos llamado `X` a las puntuaciones de los restantes animales. Calculamos primero las similaridades, según el coeficiente de Sokal y Michener, entre este nuevo individuo y los demás y también los cuadrados de las distancias asociadas:

```
[n,p] = size(X);
x = [1 1 0 0 0 1];
a = X*x'; d=(ones(n,p)-X)*(ones(1,p)-x)';
s = (a+d)/p;
d = 2*(ones(n,1)-s);
```

y obtenemos

```
s' =  0.8333    0.8333    0.6667    0.6667    0.6667    0.3333
d' =  0.3333    0.3333    0.6667    0.6667    0.6667    1.3333
```

Implementando la fórmula (5.12) como sigue

```
B = Y*Y';
b = diag(B);
[n,p] = size(Y);
Lambda = diag(vaps(1:p));
y = 1/2*inv(Lambda)*Y'*(b-d);
```

obtenemos las coordenadas del nuevo individuo:

```
y' =  0.1491    0.3582    0.0000   -0.1861
```

Para añadir este nuevo punto a la representación gráfica, hacemos:

```
hold on
plot(y(1),y(2),'*r','MarkerSize',15);
```

La Figura 5.5 contiene esta nueva representación de los animales. Observemos que el *elefante* está "donde corresponde", puesto que es un herbívoro salvaje.

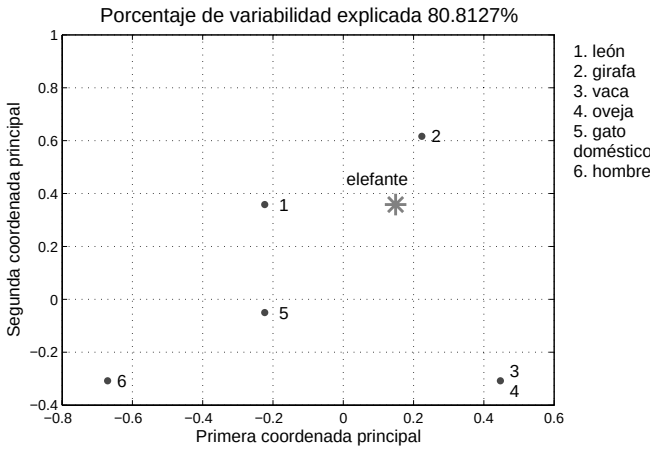


Figura 5.5.
Fórmula de interpolación de Gower (Problema 5.11).

CAPÍTULO 6

Análisis de conglomerados

Sea $\mathcal{E}$ un conjunto de n objetos o individuos sobre los que se ha calculado alguna medida de distancia. Sea $\mathbf{D} = (\delta_{ij})_{1\leq i,j\leq n}$ la matriz de distancias entre estos n individuos.

El objetivo del análisis de conglomerados (en inglés, *cluster analysis*) es la *clasificación* (no supervisada) de los elementos de $\mathcal{E}$, es decir, su agrupación en clases disjuntas, que se denominan *conglomerados* (o *clusters*). Si estas clases se agrupan sucesivamente en clases de un nivel superior, el resultado es una estructura jerárquica de conglomerados, que puede representarse gráficamente mediante un árbol, llamado *dendrograma*.

Se dice que una matriz de distancias $\mathbf{D}$ es *ultramétrica* si todos los elementos de $\mathcal{E}$ verifican la *desigualdad ultramétrica* (véase el Capítulo 5). Puede demostrarse que a cada dendrograma le corresponde una matriz de distancias ultramétrica y viceversa.

Como ocurría en el caso euclídeo, una matriz de distancias obtenida de unos datos en general no es ultramétrica. Esto da lugar al problema de aproximar la matriz de distancias $\mathbf{D}$ con una matriz ultramétrica $\mathbf{U}$ según algún criterio de proximidad adecuado.

PROBLEMA 6.1

Sea δ una función de distancia sobre los elementos de un conjunto $\mathcal{E}$ que verifica la desigualdad ultramétrica.

(a) *Sean $i, j, k \in \mathcal{E}$ tales que $\delta_{ij} = a$, $\delta_{ik} = b$, $\delta_{jk} = c$, con $a \leq b \leq c$. Demuéstrese que $b = c$.*

(b) *Usando el apartado (a) demuéstrese que δ cumple la desigualdad triangular.*

SOLUCIÓN

(a) Puesto que δ verifica la desigualdad ultramétrica y, además $a \leq b \leq c$:

$$\left.\begin{array}{l} b = \delta_{ik} \leq \max\{\delta_{ij}, \delta_{jk}\} = \max\{a, c\} = c \\ c = \delta_{jk} \leq \max\{\delta_{ji}, \delta_{ik}\} = \max\{a, b\} = b \end{array}\right\} \Rightarrow b = c.$$

Esto significa que con una distancia ultramétrica todo triángulo es isósceles.

(b) Debemos comprobar que se cumple la desigualdad $\delta_{ij} \leq \delta_{ik} + \delta_{kj}$,para todo i, j, k, teniendo en cuenta que $b = c$. Consideremos los tres posibles casos:

$$\left.\begin{array}{c} \delta_{ij} \leq \delta_{ik} + \delta_{kj} \\ \delta_{ij} = a \\ \delta_{ik} + \delta_{kj} = b + c \end{array}\right\} \Leftrightarrow a \leq b + c \quad \text{Es cierto, puesto que } a \leq b \leq c.$$

$$\left.\begin{array}{c} \delta_{ik} \leq \delta_{ij} + \delta_{jk} \\ \delta_{ik} = b \\ \delta_{ij} + \delta_{jk} = a + c = a + b \end{array}\right\} \Leftrightarrow b \leq a + b \quad \text{Es cierto, puesto que } a > 0.$$

$$\left.\begin{array}{c} \delta_{jk} \leq \delta_{ji} + \delta_{ik} \\ \delta_{jk} = c \\ \delta_{ji} + \delta_{ik} = a + b = a + c \end{array}\right\} \Leftrightarrow c \leq a + c \quad \text{Es cierto, puesto que } a > 0.$$

PROBLEMA 6.2

La Tabla 6.1 contiene las distancias por carretera (en km) entre 5 ciudades españolas. Realícese una clasificación jerárquica mediante el método del mínimo (o single linkage*). Obténgase la matriz de distancias ultramétrica.*

SOLUCIÓN

Para abreviar, denotaremos las ciudades por sus iniciales y trabajaremos solamente con el triángulo superior de la matriz de distancias. El paso cero del algoritmo de clasificación consiste en expresar la unión disjunta formada por cada uno de los elementos del conjunto, es decir, $C_0 = \{B\} + \{M\} + \{SS\} + \{S\} + \{V\}$.

Tabla 6.1.
Distancias por carretera (en km) entre ciudades. (Problema 6.2)

	Barcelona	Madrid	San Sebastián	Sevilla	Valencia
Barcelona	0	639	606	1181	364
Madrid	639	0	474	542	355
San Sebastián	606	474	0	908	597
Sevilla	1181	542	908	0	679
Valencia	364	350	597	679	0

En el primer paso del algoritmo se juntan los individuos más cercanos, que en este caso son las ciudades Madrid y Valencia, puesto que $\delta_{M,V} = 355$. Estas dos ciudades forman el primer conglomerado. De manera que en el paso 1 la clasificación será:

$$C_1 = \{B\} + \{M, V\} + \{SS\} + \{S\}.$$

Ahora mediante el método del mínimo hay que recalcular las distancias del conglomerado {M,V} a los demás individuos:

$$\delta_{(MV),B} = \min\{\delta_{M,B}, \delta_{V,B}\} = \min\{639, 364\} = 364,$$
$$\delta_{(MV),SS} = \min\{\delta_{M,SS}, \delta_{V,SS}\} = \min\{474, 597\} = 474,$$
$$\delta_{(MV),S} = \min\{\delta_{M,S}, \delta_{V,S}\} = \min\{542, 679\} = 542,$$

de manera que la matriz de distancias ha quedado:

Paso 0	B	M	SS	S	V
B	0	639	606	1181	364
M		0	474	542	**355**
SS			0	908	597
S				0	679
V					0

$\rightarrow$

Paso 1	B	(M,V)	SS	S
B	0	**364**	606	1181
(M,V)		0	474	542
SS			0	908
S				0

Se prosigue análogamente hasta que se obtenga un conglomerado que contenga a todos los individuos. El siguiente conglomerado que se forma es {B,M,V}, puesto que Barcelona es la ciudad más cercana al conglomerado {M,V} al ser $\delta_{B,MV} = 364$. En este segundo paso, la clasificación será $C_2 = \{B, M, V\} + \{SS\} + \{S\}$. Como anteriormente, hay que recalcular las distancias del conglomerado {B,M,V} al resto de individuos:

$$\delta_{(BMV),SS} = \min\{\delta_{B,SS}, \delta_{(MV),SS}\} = \min\{606, 474\} = 474,$$
$$\delta_{(BMV),S} = \min\{\delta_{B,S}, \delta_{(MV),S}\} = \min\{1181, 542\} = 542,$$

y la matriz de distancias es:

Paso 2	(B,MV)	SS	S
(B,MV)	0	**474**	542
SS		0	908
S			0

$\rightarrow$

Paso 3	(BMV,SS)	S
(BMV,SS)	0	542
S		0

En el tercer paso se forma el conglomerado {B,M,V,SS}, puesto que San Sebastián es la ciudad más próxima al conglomerado {B,M,V} con $\delta_{SS,BMV} = 474$. Ahora la clasificación es

$$C_3 = \{B, M, V, SS\} + \{S\}$$

y la distancia del conglomerado {B,M,V,SS} al individuo que falta es:

$$\delta_{(BMVSS),S} = \min\{\delta_{(BMV),S}, \delta_{SS,S}\} = \min\{542, 908\} = 542,$$

Finalmente, en el paso 4 se forma el último conglomerado {B,M,V,SS,S} con una distancia de $\delta_{BMVSS,S} = 542$, que es lo que dista Sevilla del conglomerado {B,M,V,SS}. La clasificación en este último paso es $C_4 = \{B, M, V, SS, S\}$. La Tabla 6.2 contiene un resumen de los conglomerados que se han ido formando en las distintas etapas del algoritmo de clasificación.

Tabla 6.2.
Resumen del algoritmo de clasificación (Problema 6.2).

paso	distancias	clasificación / conglomerados
0	-	$C_0 = \{B\} + \{M\} + \{SS\} + \{S\} + \{V\}$
1	$\delta_{M,V} = 355$	$C_1 = \{B\} + \{M, V\} + \{SS\} + \{S\}$
2	$\delta_{B,MV} = 364$	$C_2 = \{B, M, V\} + \{SS\} + \{S\}$
3	$\delta_{BMV,SS} = 474$	$C_3 = \{B, M, V, SS\} + \{S\}$
4	$\delta_{BMVSS,S} = 542$	$C_4 = \{B, M, V, SS, S\}$

A partir de la Tabla 6.2 puede reconstruirse la matriz de distancias ultramétrica, que se muestra en la Tabla 6.3. La representación de los individuos a partir de las distancias ultramétricas suele realizarse mediante un dendrograma o árbol jerárquico. Si la matriz de distancias originales no cumple la propiedad ultramétrica, los distintos métodos de clasificación darán lugar a distintos dendrogramas.

Tabla 6.3.
Matriz de distancias ultramétrica entre ciudades. (Problema 6.2).

	Barcelona	Madrid	San Sebastián	Sevilla	Valencia
Barcelona	0	364	474	542	364
Madrid		0	474	542	355
San Sebastián			0	542	474
Sevilla				0	542
Valencia					0

La Figura 6.1 contiene una representación, en forma de árbol jerárquico o dendrograma, de la matriz de distancias ultramétrica calculada mediante el método del mínimo.

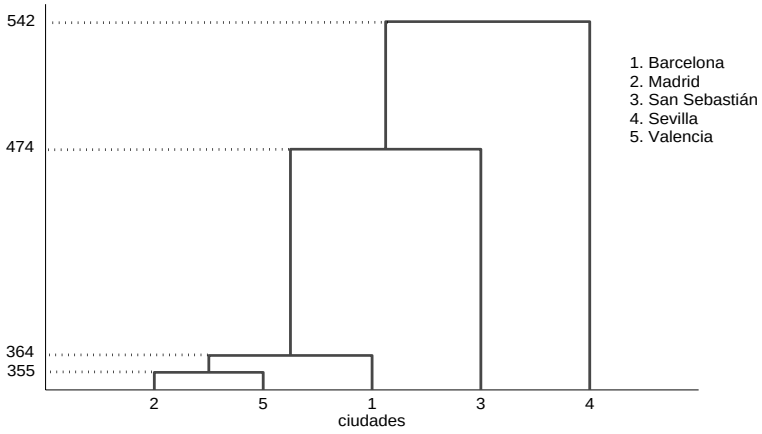


Figura 6.1.
Dendrograma con los datos del Problema 6.2

PROBLEMA 6.3

Considérense los datos de la Tabla 5.1. Sea $\mathbf{D}^{(2)}$ *la matriz de distancias de Bhattacharyya obtenida en el Problema 5.3.*

(a) *Verifíquese que* $\mathbf{D}$ *no es ultramétrica.*

(b) *Realícense clasificaciones jerárquicas a partir de la matriz* $\mathbf{D}$ *mediante los métodos del mínimo (o* single linkage*), del máximo (o* complete linkage*) y UPGMA (o* Unweighted Pair Group Method using Arithmetic averages*). ¿Qué diferencias se observan?*

(c) *Calcúlese la correlación cofenética en cada caso.*

(d) *Compárense los dendrogramas con la representación en coordenadas principales que muestra la Figura 5.2.*

SOLUCIÓN

(a) En el Problema 5.3 habíamos calculado la matriz `DB2` de cuadrados de distancias de Bhattacharyya con los datos de la Tabla 5.1. Esta tabla contenía las proporciones génicas (observadas) entre 10 poblaciones. De manera que el conjunto de individuos sobre el que queremos realizar clasificaciones jerárquicas es $\mathcal{E} = \{$*francesa, checa, germánica, vasca, china, ainu, esquimal, negra USA, espa nola, egipcia*$\}$.

En primer lugar, calculamos la matriz de distancias `D=sqrt(DB2)`, para ver si los elementos de $\mathcal{E}$ cumplen o no la propiedad ultramétrica:

```
D=[  0  0.3959  0.2086  0.3530  0.5351  0.6298  0.5121  0.4301  0.2828  0.4695
0.3959       0  0.3400  0.5162  0.4733  0.5104  0.4976  0.4575  0.3693  0.2995
0.2086  0.3400       0  0.4074  0.5211  0.6030  0.5107  0.4206  0.2789  0.4227
0.3530  0.5162  0.4074       0  0.5675  0.6879  0.5106  0.5055  0.3895  0.5796
0.5351  0.4733  0.5211  0.5675       0  0.4397  0.4354  0.5206  0.5151  0.4991
0.6298  0.5104  0.6030  0.6879  0.4397       0  0.5569  0.6084  0.6035  0.4921
0.5121  0.4976  0.5107  0.5106  0.4354  0.5569       0  0.6007  0.4499  0.5680
0.4301  0.4575  0.4206  0.5055  0.5206  0.6084  0.6007       0  0.4938  0.4469
0.2828  0.3693  0.2789  0.3895  0.5151  0.6035  0.4499  0.4938       0  0.4702
0.4695  0.2995  0.4227  0.5796  0.4991  0.4921  0.5680  0.4469  0.4702  0 ];
```

Puede comprobarse que la matriz `D` no es ultramétrica puesto que, por ejemplo,

$$\delta_{1,6} = 0.6298 > \max\{\delta_{1,3}, \delta_{3,6}\} = \max\{0.2086, 0.6030\}.$$

(b) Para poder utilizar las funciones incorporadas en Matlab que permiten realizar el análisis de conglomerados, necesitamos expresar la matriz de distancias como un vector fila que contenga solamente la parte triangular superior de la matriz, pero sin la diagonal principal. Para ello, podemos utilizar la siguiente función:

```
% la funcion Y=extractdist(D) extrae las distancias de los
% elementos de la parte triangular superior (sin contar la
% diagonal) de la matriz D (nxn) de distancias. Los elementos
% se extraen ordenadamente, columna a columna.
%
% Entradas: D  es una matriz cuadrada (nxn).
% Salidas: Y es un vector fila de dimension n(n-1)/2.
%
function Y = extractdist(D)
[n,n] = size(D);
Y = [D(1,2:n)];
for i = 2:n-1
   Y = [Y D(i,i+1:n)];
end
```

Podéis comprobar que mediante la instrucción `Y=squareform(D)` se llega al mismo resultado.
Utilizando las funciones internas de Matlab `linkage` y `dendrogram` (sólo disponibles con la Toolbox Statistics) se obtiene una representación en forma de árbol jerárquico o dendrograma. La función `linkage` da lugar a una matriz de 3 columnas, que contiene el índice de la jerarquía indexada en su tercera columna y, por tanto, permite recuperar la matriz de distancias ultramétrica, si ésta fuera de interés.

```
Z_min = linkage(Y,'single');
Z_max = linkage(Y,'complete');
Z_UPGMA = linkage(Y,'average');
dendrogram(Z_min);
dendrogram(Z_max);
dendrogram(Z_UPGMA);
```

La Figura 6.2 contiene los dendrogramas obtenidos mediante los tres métodos anteriores. Observad que las clasificaciones que se obtienen mediante los métodos del máximo y UPGMA son muy parecidas. Por otro lado, el método del mínimo tiende a contraer el espacio (observad los valores del índice de la jerarquía, que se encuentran representados en el eje vertical del gráfico), mientras que el método de máximo tiende a dilatar el espacio.

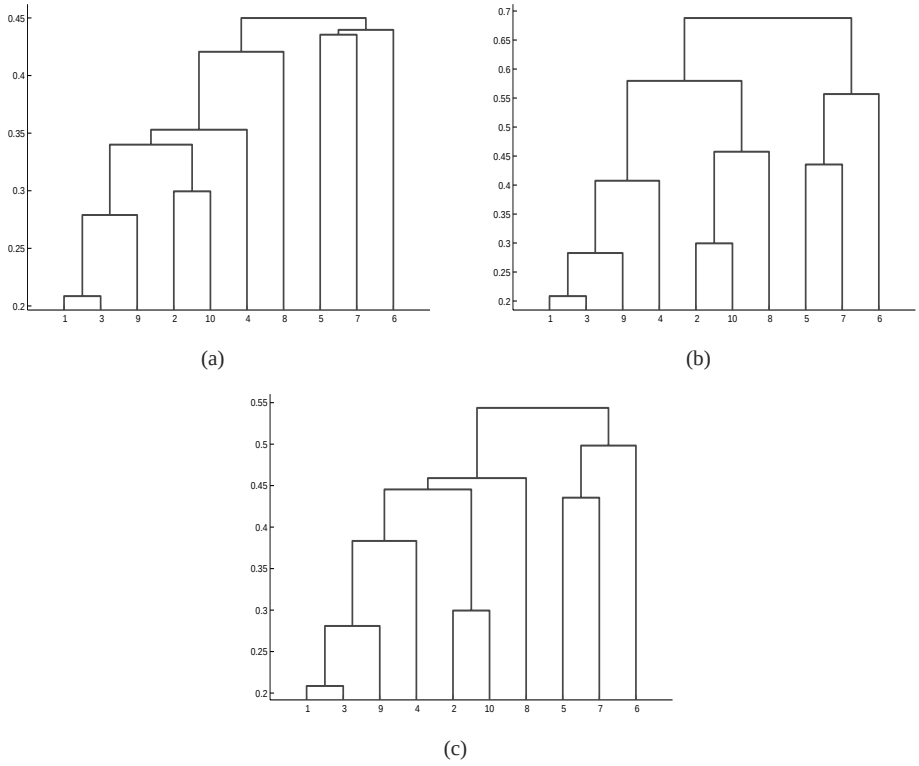


Figura 6.2.
Dendrogramas con los datos del Problema 6.3: métodos (a) del mínimo, (b) del máximo y (c) UPGMA.

(c) La correlación cofenética es el coeficiente de correlación lineal de Pearson entre los elementos de la matriz de distancias original y los elementos de la matriz de distancias ultramétrica. Se utiliza como medida de proximidad entre las dos matrices de distancias. Este coeficiente vale uno en caso de proporcionalidad (igualdad) de ambas matrices, lo que equivale a decir que la matriz de distancias original ya cumple la propiedad ultramétrica.
Para calcular la correlación cofenética podemos utilizar la función interna de Matlab `cophenet`:

```
c_min = cophenet(Z_min,Y)
c_max = cophenet(Z_max,Y)
c_UPGMA = cophenet(Z_UPGMA,Y)
```

y obtenemos `c_min=0.7910`, `c_max=0.8132` y `c_UPGMA=0.8413`, indicando que el

método UPGMA es el que menos distorsiona (de los tres que hemos visto) la matriz de distancias original. Los métodos del tipo UPGMA se utilizan mucho en biología porque maximizan la correlación cofenética.

(d) Las agrupaciones de individuos que se observan en los dendrogramas deberían reflejarse también en la representación en coordenadas principales de estos mismos individuos (Figura 5.2). La Figura 6.3 intenta reflejar estas proximidades. El grupo {1,3,9,4} lo forman las poblaciones europeas {*francesa, germánica, espa nola, vasca*}, el grupo {2,8,10} está formado por las poblaciones {*checa, negra USA, egipcia*} y, finalmente, el grupo {5,6,7} lo forman las poblaciones {*china, ainu, esquimal*}. Observad que los dendrogramas obtenidos mediante el método del máximo y mediante el método UPGMA son los más parecidos a las agrupaciones que muestra la Figura 6.3.

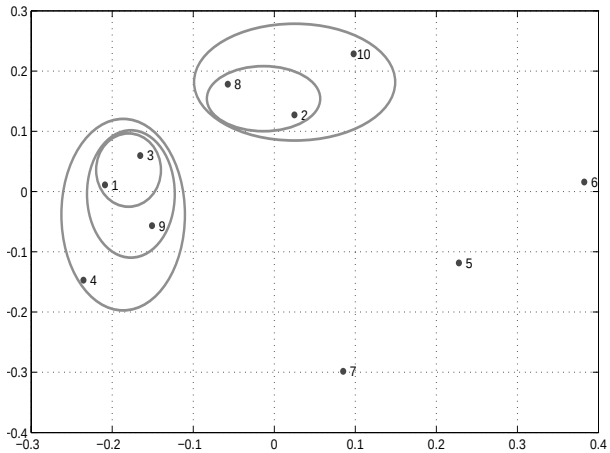


Figura 6.3.
Representación en coordenadas principales y agrupaciones (Problema 5.3)

PROBLEMA 6.4

Considérense los datos del Problema 5.4. Sea $\mathbf{D}^{(2)}$ *la matriz de cuadrados de distancias obtenida a partir del coeficiente de similaridad de Sokal y Michener.*

(a) *Verifíquese que* $\mathbf{D}$ *no es ultramétrica.*

(b) *Realícense clasificaciones jerárquicas mediante los métodos del mínimo, del máximo y UPGMA. ¿Qué diferencias se observan?*

(c) *Calcúlese la correlación cofenética en cada caso.*

(d) *Compárense los dendrogramas con la representación en coordenadas principales que muestra la Figura 5.4*

SOLUCIÓN

(a) El conjunto de individuos sobre el que debemos realizar clasificaciones jerárquicas es

$$\mathcal{E} = \{\textit{león, jirafa, vaca, oveja, gato doméstico, hombre}\}.$$

A partir de la matriz `D2_Sokal` de cuadrados de distancias obtenida en el Problema 5.4, calculamos la matriz de distancias:

```
D = sqrt(D2_Sokal);

D =[   0    0.8165    1.0000    1.0000    0.5774    1.0000
  0.8165         0    1.0000    1.0000    1.0000    1.2910
  1.0000    1.0000         0         0    0.8165    1.1547
  1.0000    1.0000         0         0    0.8165    1.1547
  0.5774    1.0000    0.8165    0.8165         0    0.8165
  1.0000    1.2910    1.1547    1.1547    0.8165         0];
```

Puede comprobarse que la matriz `D` no es ultramétrica puesto que, por ejemplo,

$$\delta_{1,3} = 1 > \max\{\delta_{1,5}, \delta_{5,3}\} = \max\{0.5774, 0.8165\}.$$

(b) Para obtener los dendrogramas haremos:

```
Y = squareform(D);
Z_min = linkage(Y,'single');
Z_max = linkage(Y,'complete');
Z_UPGMA = linkage(Y,'average');
dendrogram(Z_min)
dendrogram(Z_max)
dendrogram(Z_UPGMA)
```

La Figura 6.4 contiene los dendrogramas obtenidos mediante los tres métodos anteriores. De nuevo puede observarse que el método del mínimo contrae el espacio, mientras que el método del máximo lo dilata.

(c) Las correlaciones son

```
c_min=0.8846,
c_max=0.8556,
c_UPGMA=0.8985.
```

(d) La Figura 6.5 contiene la representación en coordenadas principales de los animales. Observad el parecido de las proximidades entre individuos que refleja esta figura con la clasificación jerárquica obtenida mediante el método del máximo.

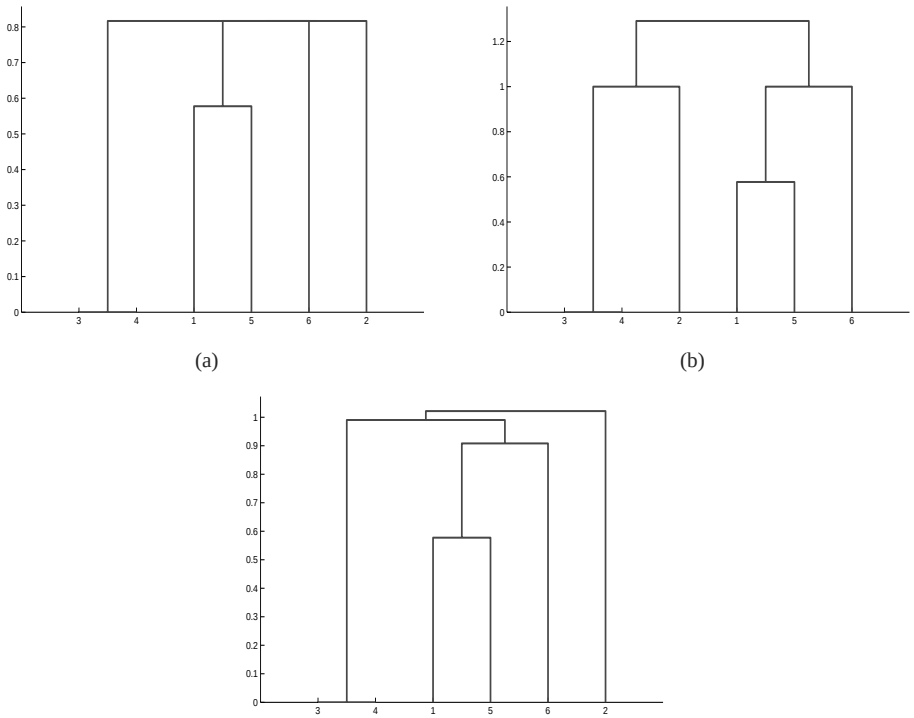


Figura 6.4.

Dendrogramas con los datos del Problema 6.4: métodos (a) del mínimo, (b) del máximo y (c) UPGMA.

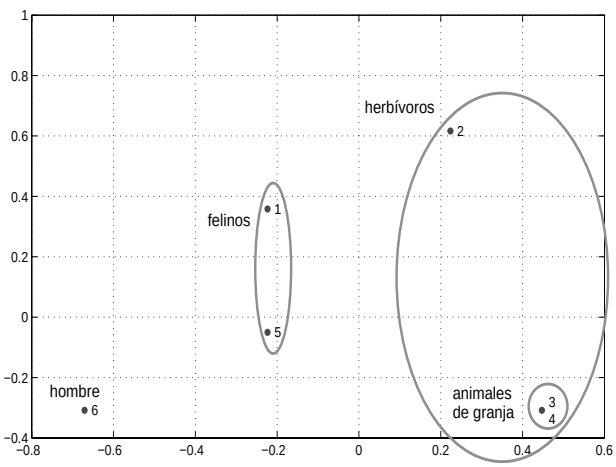


Figura 6.5.

Representación en coordenadas principales y agrupaciones (Problema 5.4)

PROBLEMA 6.5

La Tabla 4.1 contiene una serie de indicadores económicos y sociales sobre 96 países del mundo. Sea $\mathbf{Y}$ *la matriz que contiene las dos primeras componentes principales calculadas a partir de la matriz de correlaciones (véase el Problema 4.4). Obténganse las distancias euclídeas entre países a partir de* $\mathbf{Y}$ *y realícese una clasificación jerárquica mediante el método UPGMA. Coméntense los resultados obtenidos.*

SOLUCIÓN

Partimos de la matriz `X` que contiene los datos de la Tabla 4.1. En primer lugar calculamos las componentes principales (véase el Problema 4.4) y nos quedamos solamente con las dos primeras componentes calculadas a partir de la matriz de correlaciones, es decir, las dos primeras columnas de `Y2`. La función interna de Matlab `pdist` permite calcular distintas funciones de distancia a partir de matrices de datos. Para calcular la distancia euclídea haremos,

```
pdist(Y2,'euclidean')
```

o, simplemente

```
pdist(Y2)
```

puesto que ésta es la distancia que la función `pdist` calcula por defecto. Si, en cambio, quisiéramos calcular la distancia de Mahalanobis, haríamos

```
pdist(Y2,'mahalanobis')
```

El siguiente código resuelve el ejercicio:

```
[T1,Y1,acum1,T2,Y2,acum2] = comp(X);
Y2 = Y2(:,1:2);
Y = pdist(Y2,'euclidean');
Z = linkage(Y,'average');
dendrogram(Z,0,'colorthreshold',1.5)
c = cophenet(Z,Y)
```

Por defecto, la función

```
dendrogram(Z,p,'colorthreshold',t)
```

genera dendrogramas a partir de los últimos $p = 30$ conglomerados formados y asigna colores distintos a los conglomerados que se forman a un nivel (o *threshold*) menor que el valor de t. Si inicialmente tenemos más de 30 individuos, como en este ejercicio, hay que indicarle que los dibuje todos mediante la opción $p = 0$. Hemos puesto $t = 1.5$ para que asigne colores distintos a los conglomerados que se han formado a una distancia menor de 1.5. La Figura 6.6 contiene el dendrograma calculado a partir del método UPGMA. Observad las distintas agrupaciones que se forman según el nivel que se considere. El coeficiente de correlación cofenética es `c = 0.8028`.

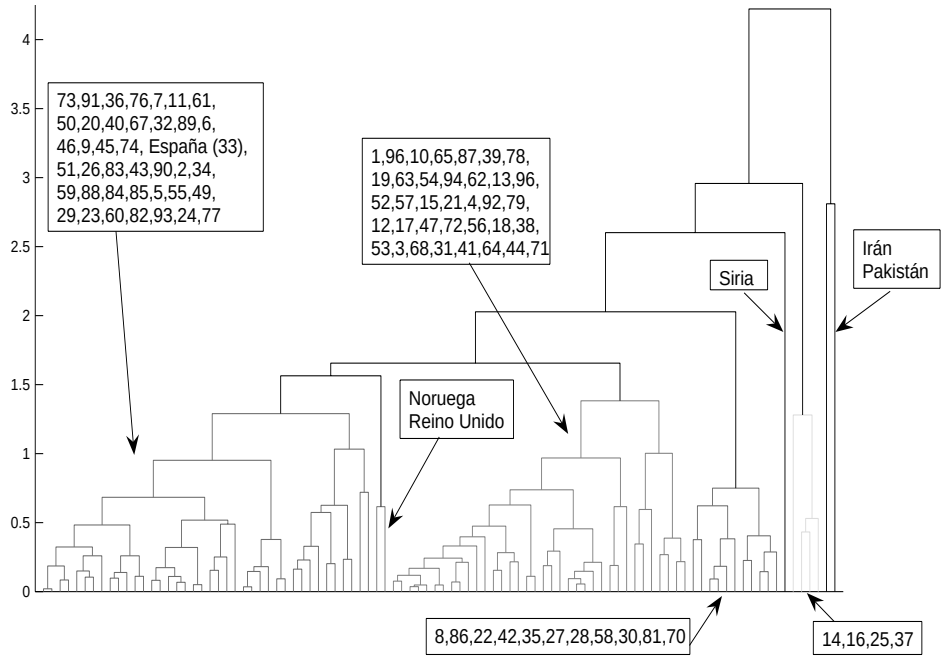


Figura 6.6.
Dendrograma con los datos del Problema 6.5.

PROBLEMA 6.6

Se ha realizado una encuesta a un grupo de personas pidiéndoles que clasificaran una lista de hortalizas según sus parecidos. La Tabla 6.4 contiene la matriz de disimilaridades entre estos hortalizas. Realícese un análisis de clasificación jerárquica mediante los métodos del centroide, de la mediana y de Ward. Obténgase la correlación cofenética en cada caso.

SOLUCIÓN

Sea `D` la matriz de disimilaridades de la Tabla 6.4. Para poder aplicar cualquiera de los tres métodos (centroide, mediana, Ward) es necesario que la matriz de disimilaridades sea euclídea. Puesto que éste no es el caso de la matriz `D`, en primer lugar debemos *euclidianizar* esta matriz de distancias. Para ello utilizaremos la función `non2euclid`, que vimos en el Problema 5.7, y que realiza este tipo de transformaciones para matrices de cuadrados de distancias.

```
D2 = D.*D;  D2_euclid = non2euclid(D2);
D_euclid = sqrt(D2_euclid);
Y = squareform(D_euclid);
```

Tabla 6.4.
Matriz de distancias entre hortalizas (Problema 6.6)

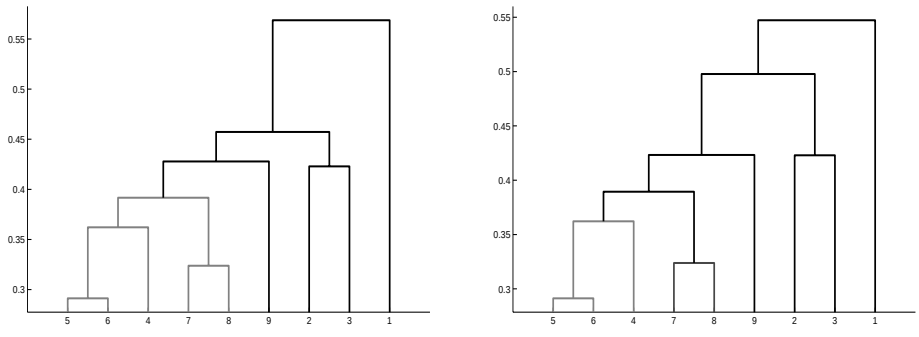
	1	2	3	4	5	6	7	8	9
1. nabo	0	0.318	0.270	0.311	0.378	0.392	0.399	0.392	0.426
2. col		0	0.101	0.223	0.243	0.236	0.311	0.345	0.358
3. remolacha			0	0.061	0.236	0.176	0.345	0.297	0.318
4. espárrago				0	0.061	0.088	0.176	0.101	0.230
5. zanahoria					0	0.007	0.074	0.209	0.264
6. espinacas						0	0.128	0.182	0.128
7. judías verdes							0	0.027	0.142
8. guisantes								0	0.128
9. maíz									0

```
Z_ward = linkage(Y,'ward');
Z_median = linkage(Y,'median');
Z_centroid = linkage(Y,'centroid');

c_ward = cophenet(Z,Y);
c_median = cophenet(Z_median,Y);
c_centroid = cophenet(Z_centroid,Y);

dendrogram(Z_ward,'colorthreshold','default')
dendrogram(Z_median,'colorthreshold','default')
dendrogram(Z_centroid,'colorthreshold','default')
```

Las Figuras 6.7 y 6.8 contienen los dendrogramas correspondientes a los tres métodos. Las correlaciones son `c_ward=0.6481`, `c_median=0.8460`, `c_centroid=0.8213`, indicando que el método de la mediana es el que menos distorsiona la aproximación de la matriz de distancias euclídeas por la de distancias ultramétricas.



(a) (b)

Figura 6.7.
Dendrogramas con los datos del Problema 6.6: métodos (a) del centroide y (b) de la mediana

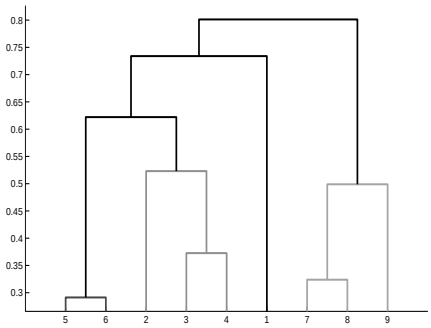


Figura 6.8.
Dendrograma con los datos del Problema 6.6: método de Ward

CAPÍTULO 7

Análisis factorial

El análisis factorial intenta describir la relación entre varias variables dependientes

$$X_1, \ldots, X_p$$

a partir de un número m, menor que p, de variables independientes y no observables, que llamaremos factores (comunes)

$$F_1, \ldots, F_m\,.$$

Algunas preguntas que se plantean en este capítulo son: ¿cómo elegir el número m de factores a utilizar?, ¿qué representan los factores comunes?, ¿cuál es el modelo que relaciona las variables originales y los factores?, ¿cuánta información proporcionan los factores comunes acerca de las variables X_i?

Existe una estrecha relación entre el análisis factorial y las componentes principales. En ambos casos se intenta aproximar la matriz de covarianzas de

$$\mathbf{X} = (X_1, \ldots, X_p)'$$

con datos de dimensión m reducida. Sin embargo, el análisis de componentes principales se centra en las varianzas de las X_i, mientras que el análisis factorial intenta explicar la estructura de correlaciones entre las variables.

PROBLEMA 7.1

Sea $\mathbf{X}$ *un vector aleatorio de dimensión* $p = 3$*, con vector de medias* $\boldsymbol{\mu} = (0, 1, 1)'$ *y matriz de varianzas-covarianzas*

$$\boldsymbol{\Sigma} = \begin{pmatrix} 3 & -4 & 2 \\ -4 & 12 & -2 \\ 2 & -2 & 3 \end{pmatrix}.$$

Se sabe que $\mathbf{X}$ *sigue un modelo factorial de un único factor, con matriz de varianzas específicas* $\boldsymbol{\Psi} = diag(1, 4, 1)$.

(a) *Escríbase el modelo factorial y calcúlese la matriz de cargas.*

(b) *Calcúlense las comunalidades y los porcentajes de variación de cada variable explicados por el factor del modelo obtenido en (a).*

(c) *Discútase si la solución a los apartados (a) y (b) es única.*

SOLUCIÓN

(a) El modelo es

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}F + \boldsymbol{\epsilon},$$

donde $\mathbf{L} = (l_{11}, l_{21}, l_{31})'$ es la matriz de cargas, F es una variable aleatoria (el factor común) con $E(F) = 0$ y $var(F) = 1$ y $\boldsymbol{\epsilon}$ es un vector aleatorio de dimensión $p = 3$ con $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $Var(\boldsymbol{\epsilon}) = \boldsymbol{\Psi}$ y $Cov(F, \boldsymbol{\epsilon}) = \mathbf{0}$. Del modelo se deduce la descomposición

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi},$$

que es equivalente a

$$\mathbf{L}\mathbf{L}' = \begin{pmatrix} l_{11} \\ l_{21} \\ l_{31} \end{pmatrix} (l_{11}, l_{21}, l_{31}) = \boldsymbol{\Sigma} - \boldsymbol{\Psi} = \begin{pmatrix} 2 & -4 & 2 \\ -4 & 8 & -2 \\ 2 & -2 & 2 \end{pmatrix}.$$

De los términos de la diagonal obtenemos $l_{11} = \pm\sqrt{2}$, $l_{21} = \mp 2\sqrt{2}$, $l_{31} = \pm\sqrt{2}$. De los términos fuera de la diagonal obtenemos signo(l_{11}) = signo(l_{31}) $\neq$ signo(l_{21}). Por tanto, $\mathbf{L} = \pm(\sqrt{2}, -2\sqrt{2}, \sqrt{2})'$ y el modelo queda

$$\begin{array}{rcl} X_1 & = & \sqrt{2}F + \epsilon_1 \\ X_2 - 1 & = & -2\sqrt{2}F + \epsilon_2 \\ X_3 - 1 & = & \sqrt{2}F + \epsilon_3 \end{array}$$

(o con los signos de F cambiados).

(b) La comunalidad h_i^2 de la variable X_i es el elemento i-ésimo de la diagonal del producto $\mathbf{L}\,\mathbf{L}'$, es decir, $h_1^2 = l_{11}^2 = 2$. Por tanto, el porcentaje de variación de X_1 explicado por F es igual a $h_1^2/V(X_1) \simeq 33.3\%$. Análogamente $h_2^2 = 8$ y el porcentaje de variación de X_2 explicado por F es un 66.6%. Y, por último, $h_3^2 = 2$ y el porcentaje de variación de X_3 explicado por F es 33.3%.

(c) En (a) ya se ha visto que la solución no es única. En general se sabe que se pueden efectuar rotaciones (que en dimensión 1 equivalen a cambiar el signo de $\mathbf{L}$). En (b) la solución sí es única.

PROBLEMA 7.2

La matriz

$$\mathbf{R} = \begin{pmatrix} 1 & 0.69 & 0.28 & 0.35 \\ & 1 & 0.255 & 0.195 \\ & & 1 & 0.61 \\ & & & 1 \end{pmatrix}$$

exhibe las correlaciones muestrales entre cuatro variables que caracterizan el estado financiero de una empresa.

(a) *Calcúlense los autovalores y autovectores de* $\mathbf{R}$.

(b) *Plantéese el modelo factorial ortogonal con* m *factores para el vector* $\mathbf{X}$ *que generó estos datos.*

(c) *Mediante el método de la componente principal, en los modelos factoriales con* $m = 2$ *y* $m = 3$ *factores, calcúlense las matrices de cargas, las comunalidades y el porcentaje que supone la comunalidad respecto a la varianza de cada variable.*

(d) *Decídase razonadamente entre el modelo con dos o tres factores.*

(e) *Para el modelo seleccionado en el apartado (d), calcúlense las correlaciones entre* Z_2 *(la variable* X_2 *estandarizada) y todos los factores. Estímese la varianza específica para* Z_2.

SOLUCIÓN

(a) Sea `R` la matriz $\mathbf{R}$ introducida en Matlab. Con la orden `eig(R)` calculamos los autovalores λ y autovectores (normalizados) $\mathbf{e}$ de esta matriz

Autovalor	Autovector
1.1139	$(-0.4243, -0.5397, 0.5123, 0.5160)'$
0.2681	$(0.6419, -0.6018, 0.2825, -0.3821)'$
2.1935	$(0.5400, 0.4938, 0.4797, 0.4842)'$
0.4245	$(0.3411, -0.3206, -0.6539, 0.5944)'$

(b) El modelo factorial ortogonal con m factores comunes $\mathbf{F} = (F_1, \ldots, F_m)'$ supone que

$$Z_i = l_{i1}F_1 + l_{i2}F_2 + \ldots + l_{im}F_m + \epsilon_i, \quad i = 1, \ldots, 4,$$

donde las Z_i son las variables X_i estandarizadas y $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_4)'$ denotan los factores específicos. Se establecen las siguientes hipótesis: $E(\mathbf{F}) = \mathbf{0}$, $Var(\mathbf{F}) = \mathbf{I}$, la matriz identidad $m \times m$, $E(\boldsymbol{\epsilon}) = \mathbf{0}$ y $Var(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \text{diag}(\psi_1, \ldots, \psi_4)$. Además $\mathbf{F}$ y $\boldsymbol{\epsilon}$ son incorrelados, es decir, $Cov(\mathbf{F}, \boldsymbol{\epsilon}) = \mathbf{0}$.

(c) Si la matriz de cargas es

$$\mathbf{L} = \begin{pmatrix} l_{11} & l_{12} & \ldots & l_{1m} \\ \vdots & \vdots & & \vdots \\ l_{41} & l_{42} & \ldots & l_{4m} \end{pmatrix},$$

el método de la componente principal en el análisis factorial con m factores proporciona la estimación

$$\mathbf{L} = [\sqrt{\lambda_1}\mathbf{e}_1, \ldots, \sqrt{\lambda_m}\mathbf{e}_m],$$

siendo $\lambda_1, \ldots, \lambda_m$ los m primeros autovalores de $\mathbf{R}$ (ordenados de mayor a menor) y siendo $\mathbf{e}_1, \ldots, \mathbf{e}_m$ los autovectores normalizados correspondientes. Concretamente, para $m = 2$:

$$\mathbf{L} = \begin{pmatrix} 0.7998 & -0.4478 \\ 0.7313 & -0.5696 \\ 0.7105 & 0.5407 \\ 0.7171 & 0.5446 \end{pmatrix}$$

Comunalidades
$h_1^2 = l_{11}^2 + l_{12}^2 = 0.8402$
$h_2^2 = 0.8593$
$h_3^2 = 0.7971$
$h_4^2 = 0.8108$

Para $m = 3$

$$\mathbf{L} = \begin{pmatrix} 0.7998 & -0.4478 & 0.2222 \\ 0.7313 & -0.5696 & -0.2089 \\ 0.7105 & 0.5407 & -0.426 \\ 0.7171 & 0.5446 & 0.3873 \end{pmatrix}$$

Comunalidades
$h_1^2 = l_{11}^2 + l_{12}^2 + l_{13}^2 = 0.8896$
$h_2^2 = 0.9029$
$h_3^2 = 0.9786$
$h_4^2 = 0.9608$

Dado que $var(Z_i) = 1$ para $i = 1, \ldots, 4$, el porcentaje que supone la comunalidad respecto a la varianza de cada Z_i coincide con la comunalidad.

(d) La varianza total en este caso es 4. El porcentaje de $VT(\mathbf{R})$ que explica el modelo con dos factores es

$$100\%(\lambda_1 + \lambda_2)/\, VT(\mathbf{R}) = 82.68\%$$

y el de tres factores es

$$100\%(\lambda_1 + \lambda_2 + \lambda_3)/\, VT(\mathbf{R}) = 93.30\%.$$

Teniendo en cuenta que, para el modelo con dos factores, h_3^2 es un poco baja es razonable quedarse con el modelo de $m = 3$ factores.

(e) Sabiendo que $Cov(Z_i, F_j) = l_{ij}$ tenemos

$$Corr(Z_2, F_1) = l_{21}/\sqrt{1 \cdot 1} = 0.7313\,,\ Corr(Z_2, F_2) = -0.5696\,,\ Corr(Z_2, F_3) = -0.2089.$$

Para el modelo con tres factores la estimación de la varianza específica de Z_2 es

$$\psi_2 = 1 - h_2^2 = 1 - 0.9029 = 0.0971\,.$$

PROBLEMA 7.3

En la Tabla 7.1 se puede ver una lista de variables que caracterizan el grado de desarrollo de algunos países del mundo. Las variables son

X_1 = *Tasa de mortalidad infantil por cada 1000 nacidos vivos,*
X_2 = *Porcentaje de mujeres en la población activa,*
X_3 = *Producto Nacional Bruto (PNB) per capita en 1995 (en $),*
X_4 = *Producción de electricidad (en millones de kw/h),*
X_5 = *Promedio de líneas telefónicas por cada 1000 habitantes,*
X_6 = *Consumo de agua per capita en* m^3 *(de 1980 a 1995),*
X_7 = *Consumo de energía per capita en 1994,*
X_8 = *Emisión de* CO_2 *per capita en 1992 (en Tm).*

(a) *Supóngase un modelo factorial ortogonal para las variables X_i estandarizadas. Utilícese el método de la componente principal para estimar la matriz de cargas en los modelos con tres y cuatro factores comunes.*

(b) *Estímense las comunalidades y las varianzas específicas para los dos modelos del apartado anterior. ¿Cuál de los dos modelos es razonable elegir?*

Tabla 7.1.
Variables socioeconómicas de algunos países (Problema 7.3)

País	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Albania	30	41	670	3903	12	94	341	1.2
Angola	124	46	410	955	6	57	89	0.5
Benín	95	48	370	6	5	26	20	0.1
Congo	90	43	680	435	8	20	331	1.6
Etiopía	112	41	100	1293	2	51	22	0.1
Ghana	73	51	390	6115	4	35	93	0.2
Haití	72	43	250	362	8	7	29	0.1
Honduras	45	30	600	2672	29	294	204	0.6
Kenia	58	46	280	3539	9	87	110	0.2
Mozambique	113	48	80	490	3	55	40	0.1
Nepal	91	40	200	927	4	150	28	0.1
Nicaragua	46	36	380	1688	23	367	300	0.6
Senegal	62	42	600	1002	10	202	97	0.4
Sudán	77	28	260	1333	3	633	66	0.1
Tanzania	82	49	120	1913	3	40	34	0.1
Yemen	100	29	260	2159	12	335	206	0.7
Zambia	109	45	400	7785	8	186	149	0.3
Zimbawe	55	44	540	7334	14	136	438	1.8

SOLUCIÓN

(a) Al definir el modelo factorial sobre las variables estandarizadas Z_i, aplicaremos el método de la componente principal sobre la matriz de correlaciones $\mathbf{R}$ cuyos cuatro mayores autovalores son

$$\lambda_1 = 3.7540\,,\ \lambda_2 = 1.9286\,,\ \lambda_3 = 0.8359\,,\ \lambda_4 = 0.7230\,.$$

La matriz de cargas obtenida mediante el método de la componente principal para tres factores es:

$$\mathbf{L} = \begin{pmatrix} -0.7235 & 0.0645 & 0.0297 \\ -0.4309 & 0.8491 & -0.0415 \\ 0.8018 & 0.2775 & 0.2472 \\ 0.4166 & 0.3978 & -0.8072 \\ 0.7958 & -0.2550 & 0.0834 \\ 0.3429 & -0.8406 & -0.2737 \\ 0.9147 & 0.2342 & 0.0129 \\ 0.8006 & 0.3764 & 0.1965 \end{pmatrix}$$

y para cuatro factores es:

$$\mathbf{L} = \begin{pmatrix} -0.7235 & 0.0645 & 0.0297 & -0.5996 \\ -0.4309 & 0.8491 & -0.0415 & 0.2041 \\ 0.8018 & 0.2775 & 0.2472 & 0.0212 \\ 0.4166 & 0.3978 & -0.8072 & -0.0331 \\ 0.7958 & -0.2550 & 0.0834 & 0.2946 \\ 0.3429 & -0.8406 & -0.2737 & -0.1754 \\ 0.9147 & 0.2342 & 0.0129 & -0.2369 \\ 0.8006 & 0.3764 & 0.1965 & -0.3830 \end{pmatrix}.$$

Observando la matriz $\mathbf{L}$ vemos que en el modelo con tres factores, F_1, F_2 y F_3, la segunda variable Z_2 y la sexta Z_6 quedarían descritas principalmente por F_2. Por otro lado, F_1 serviría para caracterizar las variables Z_1, Z_3, Z_5, Z_7 y Z_8, y, por tanto, representaría el grado de desarrollo económico e industrial del país. El tercer factor está únicamente determinado por la producción de electricidad. Observemos que los pesos de la cuarta columna de $\mathbf{L} = \{l_{ij}\}$ en el modelo con cuatro factores no son excesivamente altos salvo en el caso de l_{14}. Esto sugiere que añadir el cuarto factor no aporta demasiada información.
El siguiente código es útil para realizar estos cálculos. Llamamos `X` a la matriz de datos y `eigsort` es una función definida en el Capítulo 4:

```
v = size(X) ;
R = corrcoef(X);
[autovectores,autovalores] = eigsort(R);
proporcion=cumsum(autovalores)/trace(R);
f = 4;
% Cargas para f=4 factores comunes (metodo: componente principal)
L = autovectores(:,[1:f]).*(ones(v(2),1)...
                 *(sqrt(autovalores([1:f],:)))');
```

(b) Recordemos que las comunalidades h_i^2 son los elementos de la diagonal de $\mathbf{L}\,\mathbf{L}'$. Como las variables Z_i están estandarizadas *var*$(Z_i) = 1$ y, por tanto, la varianza específica es *diag*$(\mathbf{\Psi}) =$ *diag*$(\mathbf{\Sigma}) -$ *diag*$(\mathbf{L}\,\mathbf{L}')$. Concretamente,

$m = 3$		$m = 4$	
$h_1^2 = 0.5285$	$\psi_1 = 0.4715$	$h_1^2 = 0.8879$	$\psi_1 = 0.1121$
$h_2^2 = 0.9084$	$\psi_2 = 0.0916$	$h_2^2 = 0.9500$	$\psi_2 = 0.0500$
$h_3^2 = 0.7810$	$\psi_3 = 0.2190$	$h_3^2 = 0.7815$	$\psi_3 = 0.2185$
$h_4^2 = 0.9833$	$\psi_4 = 0.0167$	$h_4^2 = 0.9844$	$\psi_4 = 0.0156$
$h_5^2 = 0.7052$	$\psi_5 = 0.2948$	$h_5^2 = 0.7920$	$\psi_5 = 0.2080$
$h_6^2 = 0.8992$	$\psi_6 = 0.1008$	$h_6^2 = 0.9299$	$\psi_6 = 0.0701$
$h_7^2 = 0.8917$	$\psi_7 = 0.1083$	$h_7^2 = 0.9478$	$\psi_7 = 0.0522$
$h_8^2 = 0.8212$	$\psi_8 = 0.1788$	$h_8^2 = 0.9679$	$\psi_8 = 0.0321$

Dado que con tres factores la comunalidad h_1^2 es baja, sería más adecuado utilizar el modelo con cuatro factores. Con Matlab haremos:

```
comunalidad = diag(L*L');
psi = diag(R-L*L') ;
```

Recomendamos al lector que, con el código empleado en este ejercicio, construya una función Matlab que permita obtener la matriz de cargas, la proporción de variabilidad, las comunalidades y las varianzas específicas, a partir de una matriz de datos `X` y de un número de factores `f`.

PROBLEMA 7.4

Sea $\mathbf{X} = (X_1, X_2, X_3)'$ *un vector aleatorio con matriz de covarianzas*

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.63 & 0.45 \\ 0.63 & 1 & 0.35 \\ 0.45 & 0.35 & 1 \end{pmatrix}.$$

(a) *Pruébese que el modelo factorial con* $m = 1$ *es válido en este caso. Calcúlense la matriz de cargas y la de varianzas específicas.*

(b) *Si se toma* $m = 2$ *¿cuál sería la aproximación de la matriz de cargas que proporcionaría el método de la componente principal?*

SOLUCIÓN

(a) Puesto que $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi}$, donde $\mathbf{L} = (l_{11}, l_{21}, l_{31})'$, entonces

$$\begin{aligned} 1 &= l_{11}^2 + \psi_1, & 0.63 &= l_{11}l_{21}, & 0.45 &= l_{11}l_{31}, \\ & & 1 &= l_{21}^2 + \psi_2, & 0.35 &= l_{21}l_{31}, \\ & & & & 1 &= l_{31}^2 + \psi_3, \end{aligned}$$

obteniendo $l_{11} = 0.9$, $l_{21} = 0.7$, $l_{31} = 0.5$, $\psi_1 = 0.19$, $\psi_2 = 0.51$, $\psi_3 = 0.75$.

(b) Dado que los dos mayores autovalores y autovectores de $\boldsymbol{\Sigma}$ son

$$\begin{array}{ll} \lambda_1 = 1.9633 & \mathbf{e}_1 = (0.6250, 0.5932, 0.5075)' \\ \lambda_2 = 0.6795 & \mathbf{e}_2 = (0.2186, 0.4911, -0.8432)' \end{array}$$

la estimación de la matriz de cargas es:

$$\mathbf{L} = \left(\sqrt{\lambda_1}\mathbf{e}_1, \sqrt{\lambda_2}\mathbf{e}_2\right) = \begin{pmatrix} 0.8757 & 0.1802 \\ 0.8312 & 0.4048 \\ 0.7111 & -0.6951 \end{pmatrix}.$$

PROBLEMA 7.5

Un banco dispone de una muestra de 51 entidades financieras que cotizan ciertos derivados financieros cuyo valor en mercado permite estimar la probabilidad de que la empresa quiebre en el plazo de un año y, en caso de quiebra, la tasa de recuperación de la misma. Las empresas observadas también han sido analizadas por dos agencias de calificación externas, que han estimado la probabilidad de quiebra a un año basándose en auditorías realizadas. En la Tabla 7.2 se pueden ver las observaciones de las siguientes variables:

X_1 = *Nivel crediticio otorgado por el banco internamente a la entidad,*
X_2 = *Número de días que ha cotizado en mercado el derivado financiero,*
X_3 = *Probabilidad de quiebra deducida del derivado,*
X_4 = *Tasa de recuperación deducida del derivado,*
X_5 = *Probabilidad de quiebra emitida por la primera agencia externa,*
X_6 = *Probabilidad de quiebra emitida por la segunda agencia externa.*

(a) *Calcúlese la matriz de correlaciones muestrales* $\mathbf{R}$.

(b) *Efectúese un análisis factorial de* $\mathbf{R}$ *con dos factores por el método de la componente principal.*

(c) *Determínense las comunalidades y la proporción de varianza total explicada con los dos factores. Explíquese si se considera necesario aumentar el número de factores comunes.*

SOLUCIÓN

(a) La matriz de correlaciones es

$$\mathbf{R} = \begin{pmatrix} 1 & 0.2050 & -0.8038 & 0.7255 & -0.5141 & -0.5971 \\ & 1 & -0.2521 & -0.0409 & -0.4053 & -0.3580 \\ & & 1 & -0.7269 & 0.7622 & 0.8813 \\ & & & 1 & -0.4105 & -0.5404 \\ & & & & 1 & 0.9370 \\ & & & & & 1 \end{pmatrix}.$$

Tabla 7.2.
Datos de entidades financieras (Problema 7.5)

Entidad	X_1	X_2	X_3	X_4	X_5	X_6
1	7.6	630	0.00070	0.36	0.00041	0.00003
2	7.8	630	0.00056	0.39	0.00041	0.00003
3	8.1	630	0.00049	0.40	0.00041	0.00003
4	7.5	630	0.00060	0.39	0.00041	0.00026
5	7.5	630	0.00047	0.40	0.00041	0.00026
6	8.3	630	0.00055	0.40	0.00041	0.00019
7	7.4	630	0.00057	0.40	0.00042	0.00026
8	6.5	630	0.00190	0.35	0.00042	0.00037
9	8.0	630	0.00088	0.38	0.00052	0.00003
10	8.0	630	0.00049	0.39	0.00052	0.00003
11	8.7	630	0.00044	0.42	0.00052	0.00000
12	8.3	630	0.00055	0.39	0.00052	0.00019
13	8.5	630	0.00032	0.40	0.00052	0.00019
14	8.6	630	0.00043	0.40	0.00052	0.00019
15	8.6	630	0.00029	0.40	0.00000	0.00000
16	8.5	630	0.00029	0.40	0.00000	0.00000
17	8.6	630	0.00031	0.40	0.00000	0.00000
18	8.7	630	0.00027	0.40	0.00000	0.00019
19	8.5	630	0.00047	0.39	0.00020	0.00000
20	8.9	630	0.00058	0.40	0.00020	0.00000
21	8.5	630	0.00032	0.40	0.00020	0.00000
22	8.7	630	0.00035	0.40	0.00020	0.00000
23	8.6	630	0.00039	0.40	0.00020	0.00019
24	8.6	630	0.00031	0.40	0.00020	0.00019
25	9.1	630	0.00029	0.42	0.00000	0.00000
26	8.7	630	0.00023	0.40	0.00000	0.00019
27	7.8	629	0.00047	0.39	0.00041	0.00026
28	7.8	629	0.00047	0.40	0.00042	0.00003
29	6.5	629	0.00109	0.39	0.00000	0.00000
30	8.5	629	0.00029	0.40	0.00000	0.00000
31	8.5	629	0.00029	0.40	0.00000	0.00000
32	7.0	627	0.00069	0.37	0.00041	0.00003
33	7.0	627	0.00106	0.34	0.00042	0.00037
34	6.7	627	0.00191	0.33	0.00042	0.00166
35	8.6	627	0.00037	0.39	0.00020	0.00019
36	7.5	625	0.00071	0.41	0.00042	0.00037
37	6.7	624	0.00171	0.32	0.00042	0.00166
38	8.1	617	0.00048	0.39	0.00052	0.00003
39	7.7	614	0.00066	0.41	0.00041	0.00026
40	6.7	613	0.00235	0.35	0.00042	0.00166
41	8.1	612	0.00043	0.38	0.00052	0.00019
42	7.5	610	0.00046	0.41	0.00041	0.00026
43	8.4	602	0.00041	0.46	0.00000	0.00000
44	8.1	594	0.00047	0.39	0.00052	0.00019
45	7.7	593	0.00047	0.38	0.00041	0.00003
46	8.5	593	0.00057	0.41	0.00052	0.00019
47	8.7	584	0.00041	0.43	0.00000	0.00000
48	8.3	573	0.00044	0.41	0.00052	0.00019
49	5.6	573	0.00408	0.33	0.00647	0.00780
50	7.3	572	0.00066	0.39	0.00042	0.00026
51	8.0	572	0.00035	0.40	0.00052	0.00019

(b) La matriz de cargas estimada por el método de la componente principal es

$$\mathbf{L} = \begin{pmatrix} -0.8251 & 0.2780 \\ -0.3680 & -0.8062 \\ 0.9594 & -0.1061 \\ -0.7404 & 0.5406 \\ 0.8592 & 0.3115 \\ 0.9277 & 0.1802 \end{pmatrix}.$$

Observemos que el primer factor representa la calidad crediticia de la entidad, mientras que el segundo describe el comportamiento del derivado.

(c) La proporción de varianza total explicada con dos factores es de un 84.07%. Las comunalidades son $h_1^2 = 0.7581$, $h_2^2 = 0.7853$, $h_3^2 = 0.9317$, $h_4^2 = 0.8404$, $h_5^2 = 0.8353$ y $h_6^2 = 0.8931$. Dado que las comunalidades son bastante altas, en principio no sería necesario añadir un tercer factor común.

PROBLEMA 7.6

Consideremos los datos del Problema 7.5.

(a) *Represéntense los pares de cargas de la matriz* $\mathbf{L}$, (l_{i1}, l_{i2}), $i = 1, \ldots, 6$, *como si fueran puntos de un plano. Rótense los ejes de coordenadas manualmente con distintos ángulos* ϕ *y represéntense las cargas rotadas en un nuevo gráfico. Decídase qué ángulo de rotación parece más adecuado.*

(b) *Demuéstrese que la matriz*

$$\mathbf{T} = \begin{pmatrix} 0.8839 & 0.4677 \\ -0.4677 & 0.8839 \end{pmatrix}$$

es ortogonal. Utilícese esta matriz para rotar la matriz de cargas obtenida en el Problema 7.5 e interprétense los coeficientes de la matriz rotada.

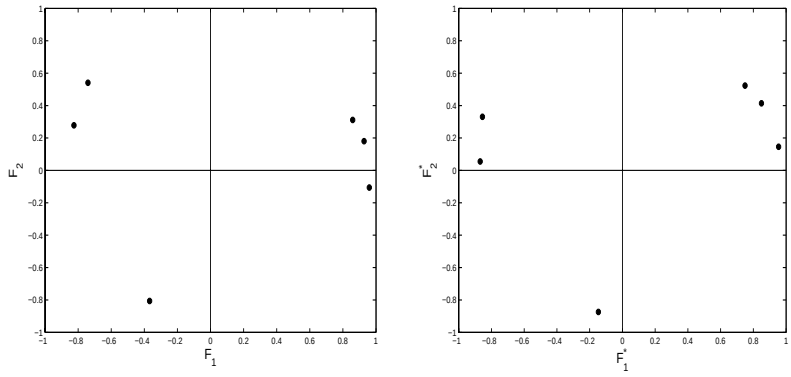
SOLUCIÓN

(a) Se puede utilizar el siguiente código (supondremos `L` ya introducida):

```
plot(L(:,1),L(:,2),'ok','MarkerFaceColor','k','MarkerSize',6)
hold on
plot([-1,1],[0,0],'-k')
hold on
plot([0,0],[-1,1],'-k')
xlabel('F_1','FontSize',16)
ylabel('F_2','FontSize',16)

phi = pi/12 ;
T = [ cos(phi) sin(phi) ; -sin(phi) cos(phi) ] ;
LRotada = L*T ;
figure(2)
plot(LRotada(:,1),LRotada(:,2),'ok','MarkerFaceColor','k',...
     'MarkerSize',6)
hold on
plot([-1,1],[0,0],'-k')
hold on
plot([0,0],[-1,1],'-k')
xlabel('F_1','FontSize',16)
ylabel('F_2','FontSize',16)
```

Observemos que la matriz T efectúa un giro en sentido contrario a las agujas del reloj. Se ha elegido ese valor de ϕ porque era el que a simple vista aproximaba más los puntos (l_{i1}, l_{i2}) a los ejes de coordenadas (véase Figura 7.1).



(a) (b)

Figura 7.1.
Representación de cargas (a) sin rotar, (b) rotadas (Problema 7.6)

(b) La matriz $\mathbf{T}$ es ortogonal porque

$$\mathbf{T}\,\mathbf{T}' = \mathbf{T}'\,\mathbf{T} = \mathbf{I}\,,$$

la matriz identidad. La matriz de cargas rotada es

$$\mathbf{L}^* = \mathbf{L}\,\mathbf{T} = \begin{pmatrix} -0.8593 & -0.1402 \\ 0.0518 & -0.8847 \\ 0.8976 & 0.3549 \\ -0.9072 & 0.1315 \\ 0.6138 & 0.6772 \\ 0.7357 & 0.5931 \end{pmatrix}.$$

Observemos que un aumento de F_1^* conlleva una disminución de la calificación interna de la entidad o de su tasa de recuperación en caso de quiebra. Por otro lado, la probabilidad de impago evaluada por cualquiera de las dos agencias crediticias es una suerte de media ponderada entre F_1^* y F_2^*, de manera que al aumentar ambos factores (por ejemplo, si disminuye el número de días de cotización del derivado), aumenta también la probabilidad de impago.

PROBLEMA 7.7

Kaiser (1958) sugirió la rotación varimax de los factores, que elige aquella matriz ortogonal $\mathbf{T}$ *que maximiza*

$$V = \frac{1}{p}\sum_{j=1}^{m}\left(\sum_{i=1}^{p}\tilde{l}_{ij}^{*4} - \frac{1}{p}\left(\sum_{i=1}^{p}\tilde{l}_{ij}^{*2}\right)^2\right),$$

donde $\mathbf{L}^* = \mathbf{L}\,\mathbf{T} = (l_{ij}^*)$ *es la matriz de las cargas rotadas,* $\tilde{l}_{ij}^* = l_{ij}^*/h_i$*, para* $i = 1, \ldots, p$*,* $j = 1, \ldots, m$*, son las cargas rotadas y reescaladas por la raíz cuadrada de la comunalidad,* m *es el número de factores y* p *es la dimensión de los datos originales. Dése una explicación intuitiva del criterio varimax. Bájese de la página web de T. Park (www.stat.ufl.edu/∼tpark/Research) el fichero* `varimaxTP.m` *y calcúlese con él la rotación varimax de la matriz* $\mathbf{L}$ *obtenida en el Problema 7.5. Represéntense en el plano las nuevas cargas rotadas.*

Indicación: *En la Statistics Toolbox de la versión 7 de Matlab y superiores ya hay programas de análisis factorial. Véase la orden* `rotatefactors` *para rotar una matriz de cargas.*

SOLUCIÓN

Observemos que

$$\frac{1}{p}\sum_{i=1}^{p}\tilde{l}_{ij}^{*4} - \left(\sum_{i=1}^{p}\frac{1}{p}\,\tilde{l}_{ij}^{*2}\right)^2 = var(\tilde{l}_j^{*2}),$$

siendo

$$\tilde{l}_j^{*2} = (\tilde{l}_{1j}^{*2}, \tilde{l}_{2j}^{*2}, \ldots, \tilde{l}_{pj}^{*2})'.$$

Por tanto,

$$V = \sum_{j=1}^{m} var(\tilde{l}_j^{*2}).$$

Maximizar V equivale a que los cuadrados de las cargas estén lo más dispersos posible sobre cada factor, de manera que las cargas sean en valor absoluto o muy grandes o muy pequeñas, pero no tomen valores intermedios.
Para calcular la rotación varimax en Matlab escribiremos

```
[RotVarimax,Lvarimax] = varimaxTP(L) ;
```

o también (si tenemos acceso a la Statistics Toolbox de Matlab 7.x)

```
[Lvarimax,RotVarimax] = rotatefactors(L,'Method','Varimax') ;
```

y dibujar las cargas rotadas `Lvarimax` como ya hicimos en el Problema 7.6.

PROBLEMA 7.8

(a) *Para los datos del Problema 7.3 y el número m de factores elegidos en el apartado (b) del mismo problema, calcúlese la rotación varimax con el programa* `varimaxTP.m` *presentado en el Problema 7.7. Calcúlese la correspondiente matriz de cargas rotada.*

(b) *Para la matriz de cargas obtenida en el apartado (a), estímense los valores observados de los factores (los llamados* scores*) por el método de mínimos cuadrados ponderados (ver, por ejemplo, Johnson y Wichern 2007).*

SOLUCIÓN

(a) En el Problema 7.3 habíamos elegido el modelo con cuatro factores. Con un código análogo al utilizado en el Problema 7.7, comprobamos que la rotación varimax viene dada por la matriz ortogonal

$$\mathbf{T} = \begin{pmatrix} 0.7439 & -0.3189 & -0.2271 & 0.5415 \\ 0.3590 & 0.8826 & -0.2885 & -0.0944 \\ 0.3124 & 0.1841 & 0.9294 & 0.0691 \\ -0.4692 & 0.2922 & 0.0379 & 0.8325 \end{pmatrix}.$$

La matriz de cargas rotada (que en Matlab llamaremos `Lvarimax`) es:

$$\mathbf{L}^* = \mathbf{L}\,\mathbf{T} = \begin{pmatrix} -0.2245 & 0.1179 & 0.1505 & -0.8950 \\ -0.1245 & 0.9389 & -0.1779 & -0.1464 \\ 0.7634 & 0.0410 & -0.0316 & 0.4428 \\ 0.2161 & 0.0600 & -0.9608 & 0.1047 \\ 0.3883 & -0.3774 & -0.0185 & 0.7060 \\ -0.0499 & -0.9529 & -0.0964 & 0.1001 \\ 0.8797 & -0.1518 & -0.2723 & 0.2769 \\ 0.9718 & 0.0011 & -0.1223 & 0.0927 \end{pmatrix}.$$

Observemos que el factor rotado F_1^* describe el comportamiento de las variables X_3 (PNB), X_7 (consumo de energía) y X_8 (emisión de CO_2), así que lo podemos interpretar como un índice del grado de desarrollo industrial del país. Los resultados F_2^* no son razonables, ya que el porcentaje de mujeres en la población activa está en relación directa con el grado de desarrollo de un país, pero el consumo de agua también. El factor F_3^* está asociado a la producción de electricidad. Por último, F_4^* describe el grado de desarrollo tecnológico y sanitario del país.

(b) Bajo la hipótesis del modelo factorial ortogonal

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\epsilon}\,,$$

con $Var(\boldsymbol{\epsilon}) = \boldsymbol{\Psi}$, y dada una muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de valores observados de $\mathbf{X}$, la estimación de los factores por mínimos cuadrados ponderados es

$$\mathbf{f}_i = (\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\boldsymbol{\Psi}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}), \quad \text{para } i = 1, \ldots, m.$$

Es habitual que los valores de $\mathbf{L}$ y $\mathbf{\Psi}$ sean desconocidos. En este caso se sustituyen por estimaciones. Cuando, por ejemplo, la unidades de medida de las X_i sean muy distintas, como es habitual se recomienda trabajar con los datos estandarizados, que es lo que haremos para este ejercicio. Para programarlo en Matlab escribimos

```
v = size(X); m = mean(X); S = cov(X,1); va = (diag(S))';
data = (X - ones(v(1),1)*m) ./ (ones(v(1),1)*va);
R = corrcoef(X);
diferencia = R - Lvarimax * Lvarimax' ;
Psi = diag(diag(diferencia)) ;
Scores = (inv(Lvarimax'*inv(Psi)*Lvarimax)*Lvarimax'...
          *inv(Psi)* data')'
```

PROBLEMA 7.9

Si suponemos que el modelo factorial ortogonal $\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\epsilon}$ *con* $Var(\epsilon) = \mathbf{\Psi}$ *es válido y que* $\mathbf{F}$ *y* $\boldsymbol{\epsilon}$ *siguen distribuciones normales, entonces* $\mathbf{X}$ *también sigue una distribución normal y es posible estimar la matriz de cargas por el método de máxima verosimilitud (véase Peña 2002, Johnson y Wichern 2007).*
Considérese la matriz de cargas

$$\mathbf{L} = \begin{pmatrix} 0.9 & 0.05 \\ 0.8 & 0.3 \\ 0.2 & 0.95 \\ 0.3 & 0.9 \\ 0.7 & 0.15 \end{pmatrix}$$

y la matriz de varianzas específicas $\mathbf{\Psi} = diag(0.2, 0.3, 0.1, 0.2, 0.3)$. *Tomando* $\boldsymbol{\mu} = \mathbf{0}$, *genérese una muestra de tamaño* $n = 1000$ *de* $\mathbf{X}$ *y obténgase la estimación de máxima verosimilitud de la matriz de cargas para* $m = 2$ *factores. Calcúlese la correspondiente estimación de la matriz de varianzas específicas.*
Indicación: *Este ejercicio sólo se puede resolver con la Statistics Toolbox de Matlab 7.x, porque incorpora la función interna* `factoran`, *que calcula la estimación de máxima verosimilitud de las cargas.*

SOLUCIÓN

Supondremos $\mathbf{L}$ y $\mathbf{\Psi}$ ya introducidas en Matlab como `L` y `Psi`. El siguiente código permite resolver el ejercicio

```
[p,m] = size(L) ;
n = 1000 ;
RaizPsi = sqrt(Psi) ;
MuestraF = randn(n,m) ;
MuestraEpsilon = randn(n,p)*RaizPsi ;
MuestraX = MuestraF * L' + MuestraEpsilon ;

L_est = factoran(MuestraX,m) ;
Psi_est = diag(diag(cov(MuestraX,1) - L_est*L_est')) ;
```

CAPÍTULO 8

Análisis canónico de poblaciones (MANOVA)

El objetivo del análisis canónico de poblaciones, o análisis multivariante de la varianza, es representar g grupos de individuos de forma óptima a lo largo de unos *ejes canónicos* ortogonales, de manera que la dispersión entre estos grupos sea máxima con relación a la dispersión dentro de los grupos. En esta representación, la distancia euclídea entre dos individuos expresados en función de los nuevos ejes canónicos coincide con la distancia de Mahalanobis entre estos individuos expresados en función de las variables originales.

Para poder aplicar correctamente esta técnica del análisis multivariante, previamente deben realizarse dos contrastes de hipótesis vistos en el Capítulo 3: *el contraste de comparación de medias*, que debe rechazarse, y *el contraste de comparación de covarianzas*, que no debe rechazarse. El hecho de inferir que las medias son iguales significa que no hay diferencias significativas entre los distintos grupos y, por tanto, la representación canónica se reduce a un solo punto. Inferir que las covarianzas no son iguales significa que los elipsoides de concentración de los distintos grupos están orientados de forma distinta y, por tanto, no se pueden determinar unos ejes comunes de representación. La hipótesis de igualdad de covarianzas raramente se cumple en las aplicaciones. A pesar de ello, si los signos de los elementos de las matrices de covarianzas muestrales de cada grupo no cambian de un grupo a otro, la orientación de los elipsoides no es demasiado distinta y todavía es posible realizar este análisis.

PROBLEMA 8.1

Considérense los datos de la Tabla 3.1. Sean $\mathbf{m}_X$, $\mathbf{m}_Y$, $\mathbf{S}_X$, $\mathbf{S}_Y$ *los vectores de medias y matrices de covarianzas correspondientes a estos datos, que se calcularon en el Problema 3.18. Sabiendo que* $n_X = 21$ *y* $n_Y = 28$,

(a) *Constrúyanse las matrices de dispersión dentro de los grupos,* $\mathbf{W}$, *y de dispersión entre los grupos,* $\mathbf{B}$.

(b) *Encuéntrese el primer eje canónico y estandarícese este eje respecto de la matriz de covarianzas común.*

(c) *Obténganse las coordenadas de los individuos medios en función del primer eje canónico estandarizado.*

(d) *Compruébese que la distancia euclídea entre los individuos medios expresados en las coordenadas canónicas coincide con la distancia de Mahalanobis entre los individuos medios expresados en las variables originales.*

SOLUCIÓN

(a) La matriz de dispersión dentro de los grupos es

$$\mathbf{W} = n_X\,\mathbf{S}_X + n_Y\,\mathbf{S}_Y = 10^3 \begin{pmatrix} 0.6278 & 0.6461 & 0.0917 & 0.0645 & 0.1049 \\ 0.6461 & 1.2289 & 0.1299 & 0.1059 & 0.1274 \\ 0.0917 & 0.1299 & 0.0303 & 0.0165 & 0.0199 \\ 0.0645 & 0.1059 & 0.0165 & 0.0152 & 0.0163 \\ 0.1049 & 0.1274 & 0.0199 & 0.0163 & 0.0472 \end{pmatrix}.$$

Sea $\mathbf{m}$ el vector de medias global, o centroide, es decir,

$$\mathbf{m} = (n_X\,\mathbf{m}_X + n_Y\,\mathbf{m}_Y)/(n_X + n_Y).$$

La matriz de dispersión entre los grupos es

$$\begin{aligned} \mathbf{B} &= n_X\,(\mathbf{m}_X - \mathbf{m})\,(\mathbf{m}_X - \mathbf{m})' + n_Y\,(\mathbf{m}_Y - \mathbf{m})\,(\mathbf{m}_Y - \mathbf{m})' \\ &= \begin{pmatrix} 13.1696 & 7.1832 & 0.5695 & -0.6738 & 0.3746 \\ 7.1832 & 3.9180 & 0.3106 & -0.3675 & 0.2043 \\ 0.5695 & 0.3106 & 0.0246 & -0.0291 & 0.0162 \\ -0.6738 & -0.3675 & -0.0291 & 0.0345 & -0.0192 \\ 0.3746 & 0.2043 & 0.0162 & -0.0192 & 0.0107 \end{pmatrix}. \end{aligned}$$

(b) La matriz de covarianzas común es

$$\mathbf{S} = \frac{1}{n_X + n_Y - 2}\,\mathbf{W} = \begin{pmatrix} 13.3576 & 13.7477 & 1.9509 & 1.3733 & 2.2309 \\ 13.7477 & 26.1459 & 2.7647 & 2.2523 & 2.7100 \\ 1.9509 & 2.7647 & 0.6445 & 0.3502 & 0.4232 \\ 1.3733 & 2.2523 & 0.3502 & 0.3244 & 0.3470 \\ 2.2309 & 2.7100 & 0.4232 & 0.3470 & 1.0035 \end{pmatrix}$$

y los ejes canónicos se obtienen a partir de la diagonalización de $\mathbf{B}$ respecto de $\mathbf{S}$. En este caso obtendremos solamente un eje canónico, puesto que sólo hay un valor propio relativo no nulo, y podemos utilizar Matlab para ello:

```
[V,L] = eig(B,S);

diag(L) = 2.8248   -0.0000   0.0000  -0.0000   0.0000
V(:,1)' = -0.3201  -0.0546  -0.1924   2.1298   0.1426
```

Si llamamos $\mathbf{v}$ a este eje canónico, puede comprobarse que ya está estandarizado respecto de $\mathbf{S}$, es decir que

$$\mathbf{v}'\,\mathbf{S}\,\mathbf{v} = 1\,.$$

(c) Las coordenadas de los individuos medios en función del primer eje canónico estandarizado son

$$\widetilde{\mathbf{m}}_X = \mathbf{m}_X'\,\mathbf{v} = -27.2238, \quad \widetilde{\mathbf{m}}_Y = \mathbf{m}_Y'\,\mathbf{v} = -27.7090.$$

(d) La distancia de Mahalanobis entre los individuos medios es

$$(\mathbf{m}_X - \mathbf{m}_Y)'\,\mathbf{S}^{-1}\,(\mathbf{m}_X - \mathbf{m}_Y) = 0.2354,$$

y la distancia euclídea entre los individuos medios en función de las coordenadas canónicas es

$$(\widetilde{\mathbf{m}}_X - \widetilde{\mathbf{m}}_Y)^2 = 0.2354\,.$$

PROBLEMA 8.2

La Tabla 8.1 contiene cuatro variables numéricas:

X_1 =*longitud del sépalo,*
X_2 =*anchura del sépalo,*
X_3 =*longitud del pétalo,*
X_4 =*anchura del pétalo,*

medidas sobre tres especies de flores del género Iris: Iris setosa, Iris versicolor *e* Iris virginica *(Fuente: Fisher 1936).*

(a) *Realícese la representación canónica de las tres especies, especificando los porcentajes de variabilidad explicados por cada eje canónico.*

(b) *Suponiendo normalidad multivariante, constrúyanse las regiones confidenciales para los individuos medios de cada grupo.*

Tabla 8.1.
Datos del Problema 8.2 (Fuente: Fisher 1936)

X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

SOLUCIÓN

Para resolver este problema utilizaremos las funciones `canp.m`, que permite obtener la representación de g grupos de individuos en unos ejes canónicos, y `regconf.m`, que permite obtener regiones confidenciales al $(1-\alpha)\,100\%$ para los individuos medios de cada grupo, suponiendo normalidad multivariante.

```
% CANP
%
% La funcion [mY,V,B,W,percent,Test1,texto1,Test2,texto2]=canp(X,n)
% realiza el analisis canonico de g poblaciones, es decir,
% representa las g poblaciones de forma optima a lo largo de
% unos ejes canonicos ortogonales.
% Para cada poblacion i (i=1,2,...,g) se tienen las medidas de
```

```
% p variables X1,X2,...,Xp sobre n(i) individuos,
% con n(1)+n(2)+...+n(g)=N.
%
% Entradas:
% X: es una matriz (N,p) que contiene las observaciones de p
%    variables (en columna) sobre los individuos de g poblaciones
%    (en fila),
% n: es un vector que contiene el numero de individuos de cada
%    poblacion.
%
% Salidas:
% mY: matriz que contiene las nuevas coordenadas de los
%     individuos medios (en fila),
% V:  matriz de vectores propios de B respecto de W (en columna),
%     es decir, las columnas de V definen los ejes canonicos,
% B:  matriz de dispersion entre poblaciones (between),
% W:  matriz de dispersion dentro de cada poblacion (within),
% percent: porcentaje de variabilidad explicado,
% Test1: vector que contiene el valor de la F de Fisher, sus
%     grados de libertad y el p-valor [F(n1,n2) n1 n2 p_valor1]
%     obtenidos en el test de comparacion de medias,
% texto1: texto resumen del resultado de Test1,
% Test2: vector que contiene el valor de la chi-cuadrado, sus
%     grados de libertad y el p-valor [chi(q) q p_valor2]
%     obtenidos en el test de comparacion de covarianzas,
% texto2: texto resumen del resultado de Test2.
%
 function [mY,V,B,W,percent,Test1,texto1,Test2,texto2]=canp(X,n)
 [N,p] = size(X);
 g = length(n);
% vector de etiquetas para las poblaciones
 for i = 1:g
   lab(i,:) = sprintf('%3g',i);
 end
%
 n0(1) = n(1);
 for i = 2:g
   n0(i) = n0(i-1)+n(i);
 end
%
% calculo de los individuos medios
%
 mX(1,:) = ones(1,n(1))*X(1:n0(1),:)/n(1);
 for i = 2:g
   mX(i,:) = ones(1,n(i))*X(n0(i-1)+1:n0(i),:)/n(i);
 end
%
% calculo de la matriz de dispersion dentro de cada poblacion
%
 H1 = eye(n(1))-ones(n(1))/n(1);
 W = X(1:n0(1),:)'*H1*X(1:n0(1),:);
 logH1 = n(1)*log(det(W/n(1)));
 for i = 2:g
   Hi = eye(n(i))-ones(n(i))/n(i);
   Ci = X(n0(i-1)+1:n0(i),:)'*Hi*X(n0(i-1)+1:n0(i),:);
   W = W+Ci;
   logH1 = logH1+n(i)*log(det(Ci/n(i)));
 end
 S = W/(N-g);
```

```
%
% calculo de la matriz de dispersion entre poblaciones
%
 mmX0 = n(1)*mX(1,:);
 for i = 2:g
    mmX = mmX0+n(i)*mX(i,:);
    mmX0 = mmX;
 end
 mmX = mmX/N;
 B0 = n(1)*(mX(1,:)-mmX)'*(mX(1,:)-mmX);
 for i = 2:g
    B = B0+n(i)*(mX(i,:)-mmX)'*(mX(i,:)-mmX);
    B0 = B;
 end
%
% Test de comparacion de medias (Lambda de Wilks).
% Conviene rechazar esta hipotesis.
%
 lambda = det(W)/det(W+B);
 [Fmit,n1,n2] = wilkstof(lambda,p,N-g,g-1);
 p_valor1 = 1-fcdf(Fmit,n1,n2);
 Test1 = [Fmit n1 n2 p_valor1];
 texto1 = char('Test1: Igualdad de medias (Lambda de Wilks): ...
                                  p-valor=',num2str(p_valor1));
%
% Test de comparacion de covarianzas (Razon de verosimilitudes
% sin la correcion de Box). Conviene aceptar esta hipotesis.
%
 logH0 = N*log(det(W/N));
 chi = logH0-logH1;
 q = (g-1)*p*(p+1)/2;
 p_valor2 = 1-chi2cdf(chi,q);
 Test2 = [chi q p_valor2];
 texto2 = char('Test2: Igualdad de covarianzas (test de Bartlett): ...
                                      p-valor=',num2str(p_valor2));
%
% calculo de los ejes canonicos
%
 [V,D] = eig(B,S);
 [z,i] = sort(-diag(real(D)));
 d = -z;
 V = real(V(:,i));
 m = min(g-1,p);
 V = V(:,1:m);
%
% estandarizacion de los ejes canonicos V'*S*V=Id.
 V = V*inv(diag(sqrt(diag(V'*S*V))));
%
% variablidad explicada
%
 for i = 1:m
   percent(i) = d(i)/sum(d)*100;
   acum(i) = sum(percent(1:i));
 end
%
% primeras dos coordenadas de los individuos
% y de los individuos medios en los nuevos ejes
%
 Y = X*V(:,1:2);
```

```
 mY = mX*V(:,1:2);
%
% representacion de los individuos y de los individuos medios
%
 if m>=2
   plot(Y(:,1),Y(:,2),'.b','MarkerSize',15)
   hold on
   plot(mY(:,1),mY(:,2),'^r','MarkerFaceColor',[1 0 0])
   grid
   xlabel('1er. eje canonico','FontSize',10)
   ylabel('2o. eje canonico','FontSize',10)
%
   title(['Coordenadas canonicas (',num2str(acum(2)),'%)'],'FontSize',12)
   for i = 1:g
      text(mY(i,1),mY(i,2),lab(i,:));
   end
 end



% REGCONF
%
% La funcion r=regconf(mY,n,p,conf) dibuja las regiones
% confidenciales para los individuos medios de g poblaciones
% obtenidos a traves de la funcion CANP.
% En cada poblacion se miden p variables sobre n(i) individuos
% (i=1,2,...,g) con n(1)+n(2)+...+n(g)=N.
%
% Entradas:
%   mY = las coordenadas canonicas de los individuos medios,
%   n =  vector columna que contiene el numero de individuos
%        de cada poblacion,
%   p = numero de variables medidas sobre cada poblacion,
%   conf = nivel de confianza (0<=conf<=1) para el que
%        se construyen las regiones confidenciales
%        (por ejemplo, conf=0.90).
%
% Salidas:
%   r = vector que contiene los radios de las esferas.
%
 function r = regconf(mY,n,p,conf)
 g = length(n);
 N = sum(n);
% valor critico de una F(p,N-g-p+1) para el nivel de
% confianza (conf) especificado.
 F = finv(conf,p,N-g-p+1);
%
% calculo de las regiones confidenciales (al conf*100%)
% para los individuos medios.
%
 for i = 1:g
   r(i) = sqrt(F*p*(N-g)/((N-g-p+1)*n(i)));
 end
 for i = 0:0.01:2*pi
   theta(floor(i*100+1)) = i;
 end
%
% vector de etiquetas para los individuos medios
%
 for i = 1:g
```

```
    lab(i,:) = sprintf('%3g',i);
  end
%
  hold on
  plot(mY(:,1),mY(:,2),'^r','MarkerFaceColor',[1 0 0])
  xlabel('1er. eje canonico','FontSize',10)
  ylabel('2o. eje canonico','FontSize',10)
%
  for i = 1:g
    for j = 1:length(theta)
       cercle(j,1) = mY(i,1)+cos(theta(j))*r(i);
       cercle(j,2) = mY(i,2)+sin(theta(j))*r(i);
    end
    plot(cercle(:,1),cercle(:,2),'.m','MarkerSize',4)
  end
  pconf = conf*100;
  title(['Regiones confidenciales para los individuos medios al ', ...
                                    num2str(pconf),'%'],'FontSize',12)
  for i = 1:g
    text(mY(i,1),mY(i,2),lab(i,:));
  end
  hold off
```

(a) Para poder utilizar la función `canp.m` debemos escribir los datos de la Tabla 8.1 en una matriz `X=[X1;X2;X3]` de dimensión $N \times p$, donde p es el número de variables observadas, y N es el número total de individuos. En este caso $p = 4$ y $N = 150$. Las matrices `X1`, `X2` y `X3` contienen a los individuos de cada uno de los tres grupos. Debemos introducir también un vector `n=[n1 n2 n3]` que contenga el número de individuos de cada grupo. Consideramos como grupo 1 a la especie *Iris setosa*, como grupo 2 a la especie *Iris versicolor* y como grupo 3 a la especie *Iris virginica*.

```
n = [50 50 50];
[mY,V,B,W,percent,Test1,texto1,Test2,texto2] = canp(X,n)
```

La matriz `mY` contiene las coordenadas de los individuos medios en función de los nuevos ejes canónicos. Las columnas de la matriz `V` son los coeficientes que definen los nuevos ejes canónicos, `B` y `W` son las matrices de dispersón entre grupos (*between*) y de dispersión dentro de los grupos (*within*), respectivamente. El vector `percent` contiene el porcentaje de variabilidad explicado por cada uno de los ejes. Puesto que el número de ejes canónicos es $\min(g-1, p)$, donde g es el número de grupos, en este caso la representación en dimensión 2 expresa el 100% de la variabilidad explicada. `Test1` y `texto1` contienen los resultados del contraste de igualdad de medias basado en el estadístico Lambda de Wilks:

```
percent = 99.1213    0.8787
Test1= 199.1453    8.0000  288.0000    0
texto1 = Test1: Igualdad de medias (Lambda de Wilks): p-valor=0
```

El primer eje canónico explica el 99.1213% de la variabilidad, mientras que el segundo eje explica solamente el 0.8787%. Para el contraste de comparación de medias se obtiene una $F(8, 288) = 199.1453$, con un p-valor asociado de 0. Por tanto se rechaza la hipótesis nula de igualdad de medias. La Figura 8.1 muestra la representación canónica de las tres especies del género Iris, con un porcentaje de variabilidad explicado del 100%.

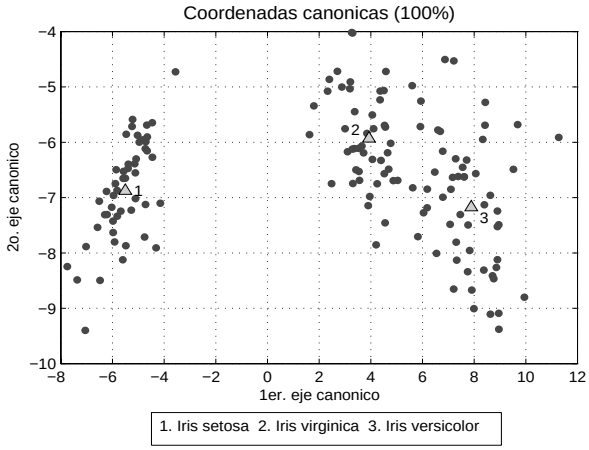


Figura 8.1.
Análisis canónico de poblaciones. (Problema 8.2.)

(b) Bajo el supuesto de normalidad multivariante, las regiones confidenciales son esferas multidimensionales centradas en los individuos medios. En el caso de la representación en dos dimensiones, se trata de círculos de radio

$$r_i = \sqrt{F_\alpha \, \frac{(N-g)\,p}{(N-g-p+1)\,n_i}}, \quad \text{para } i = 1, 2, \ldots, g,$$

donde F_α es el percentil $(1-\alpha)\,100\%$ de la ley F de Fisher con p y $N-g-p+1$ grados de libertad, p es el número de variables observadas, g es el número de grupos, N es el número total de individuos y n_i es el número de individuos en el grupo i-ésimo. Para representar las regiones confidenciales al $(1-\alpha)\,100\%$ para los individuos medios utilizaremos la función `regconf.m`. Por ejemplo, para un nivel de confianza del 90%, obtenemos:

```
r = regconf(mY,n,4,0.90)
r =  0.4026    0.4026    0.4026
```

La coincidencia de los tres radios se debe a que los tres grupos tienen el mismo número de individuos. La Figura 8.2 muestra la representación canónica de las tres especies del género Iris junto con las regiones confidenciales para los individuos medios. A veces, cuando el número de individuos es muy grande o también cuando el número de grupos es considerable, suele realizarse solamente una representación de los individuos medios juntamente con las regiones confidenciales.

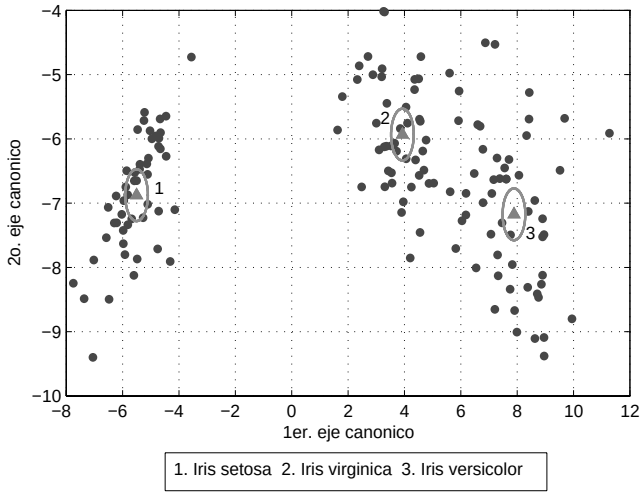


Figura 8.2.
Regiones confidenciales al 90% (Problema 8.2)

PROBLEMA 8.3

*La Tabla 8.2 contiene cuatro medidas sobre cráneos de varones egipcios de cinco períodos históricos distintos (*Grupo 1*: 4000 aC,* Grupo 2*: 3300 aC,* Grupo 3*: 1850 aC,* Grupo 4*: 200 aC,* Grupo 5*: 150 dC). Para cada período temporal se midieron 30 cráneos. Las variables observadas son: X_1 =anchura máxima, X_2 =altura basibregmática, X_3 =longitud basialveolar, X_4 =longitud de la nariz. Estos datos están accesibles en la página web* DASL Project *(véase Hutcheson y Meyer 1996).*

(a) *Realícese la representación canónica de los cinco grupos, especificando los porcentajes de variabilidad explicados por los ejes canónicos.*

(b) *Represéntense las regiones confidenciales para un nivel de confianza del* 90%.

(c) *Interprétese el primer eje canónico.*

(d) *Obténgase la matriz de distancias entre los cinco grupos.*

SOLUCIÓN

(a) Sea `X=[X1;X2;X3;X4;X5]` la matriz 150×4 que contiene los datos de la Tabla 8.2. Realizaremos el análisis canónico de poblaciones mediante la función `canp.m`:

```
n = [30 30 30 30 30];
[mY,V,B,W,percent,Test1,texto1,Test2,texto2] = canp(X,n)
```

Tabla 8.2.

Datos del Problema 8.3 (http://lib.stat.cmu.edu/DASL/Datafiles/EgyptianSkulls.html)

4000 a.C.				3300 a.C.				1850 a.C.				200 a.C.				150 d.C.			
X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
131	138	89	49	124	138	101	48	137	141	96	52	137	134	107	54	137	123	91	50
125	131	92	48	133	134	97	48	129	133	93	47	141	128	95	53	136	131	95	49
131	132	99	50	138	134	98	45	132	138	87	48	141	130	87	49	128	126	91	57
119	132	96	44	148	129	104	51	130	134	106	50	135	131	99	51	130	134	92	52
136	143	100	54	126	124	95	45	134	134	96	45	133	120	91	46	138	127	86	47
138	137	89	56	135	136	98	52	140	133	98	50	131	135	90	50	126	138	101	52
139	130	108	48	132	145	100	54	138	138	95	47	140	137	94	60	136	138	97	58
125	136	93	48	133	130	102	48	136	145	99	55	139	130	90	48	126	126	92	45
131	134	102	51	131	134	96	50	136	131	92	46	140	134	90	51	132	132	99	55
134	134	99	51	133	125	94	46	126	136	95	56	138	140	100	52	139	135	92	54
129	138	95	50	133	136	103	53	137	129	100	53	132	133	90	53	143	120	95	51
134	121	95	53	131	139	98	51	137	139	97	50	134	134	97	54	141	136	101	54
126	129	109	51	131	136	99	56	136	126	101	50	135	135	99	50	135	135	95	56
132	136	100	50	138	134	98	49	137	133	90	49	133	136	95	52	137	134	93	53
141	140	100	51	130	136	104	53	129	142	104	47	136	130	99	55	142	135	96	52
131	134	97	54	131	128	98	45	135	138	102	55	134	137	93	52	139	134	95	47
135	137	103	50	138	129	107	53	129	135	92	50	131	141	99	55	138	125	99	51
132	133	93	53	123	131	101	51	134	125	90	60	129	135	95	47	137	135	96	54
139	136	96	50	130	129	105	47	138	134	96	51	136	128	93	54	133	125	92	50
132	131	101	49	134	130	93	54	136	135	94	53	131	125	88	48	145	129	89	47
126	133	102	51	137	136	106	49	132	130	91	52	139	130	94	53	138	136	92	46
135	135	103	47	126	131	100	48	133	131	100	50	144	124	86	50	131	129	97	44
134	124	93	53	135	136	97	52	138	137	94	51	141	131	97	53	143	126	88	54
128	134	103	50	129	126	91	50	130	127	99	45	130	131	98	53	134	124	91	55
130	130	104	49	134	139	101	49	136	133	91	49	133	128	92	51	132	127	97	52
138	135	100	55	131	134	90	53	134	123	95	52	138	126	97	54	137	125	85	57
128	132	93	53	132	130	104	50	136	137	101	54	131	142	95	53	129	128	81	52
127	129	106	48	130	132	93	52	133	131	96	49	136	138	94	55	140	135	103	48
131	136	114	54	135	132	98	54	138	133	100	55	132	136	92	52	147	129	87	48
124	138	101	46	130	128	101	51	138	133	91	46	135	130	100	51	136	133	97	51

El vector `percent` contiene los porcentajes de variabilidad explicados por los 4 ejes canónicos. Se rechaza la comparación de medias con un p-valor menor que 10^{-6} y no se rechaza la igualdad de covarianzas, puesto que el p-valor asociado es de 0.12905. Estos resultados confirman que el análisis canónico de poblaciones es aplicable para este conjunto de datos.

```
percent = 88.2272    8.0941    3.2594    0.4193
Test1 =  3.8968   16.0000  434.0000    0.0000
texto1 = Test1: Igualdad de medias (Lambda de Wilks): p-valor=7.1776e-007
Test2 = 50.2206   40.0000    0.1291
texto2 = Test2: Igualdad de covarianzas (test Bartlett): p-valor=0.12905
```

La Figura 8.3 muestra la representación canónica de los individuos en dos dimensiones con un 96.32% de la variabilidad explicada.

(b) Para representar las regiones confidenciales utilizaremos la función `regconf.m`:

```
r = regconf(mY,n,4,0.90)
r = 0.5198    0.5198    0.5198    0.5198    0.5198
```

La Figura 8.4 contiene la representación canónica de los individuos junto con las regiones confidenciales para los individuos medios.

(c) Los ejes canónicos son las columnas de la matriz `V`, de manera que las coordenadas canónicas son combinaciones lineales de las variables originales, es decir, si `Y` es la representación de los individuos en las nuevas coordenadas, `Y=XV`. Así, el primer eje canónico es

$$Y_1 = 0.1267\, X_1 - 0.0370\, X_2 - 0.1451\, X_3 + 0.0829\, X_4.$$

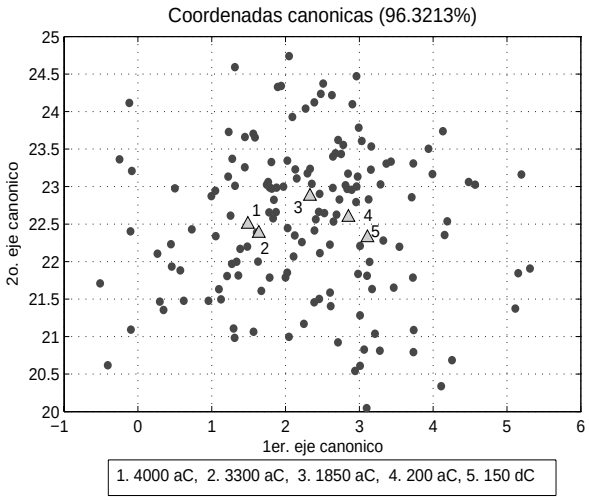


Figura 8.3.
Análisis canónico de poblaciones (Problema 8.3)

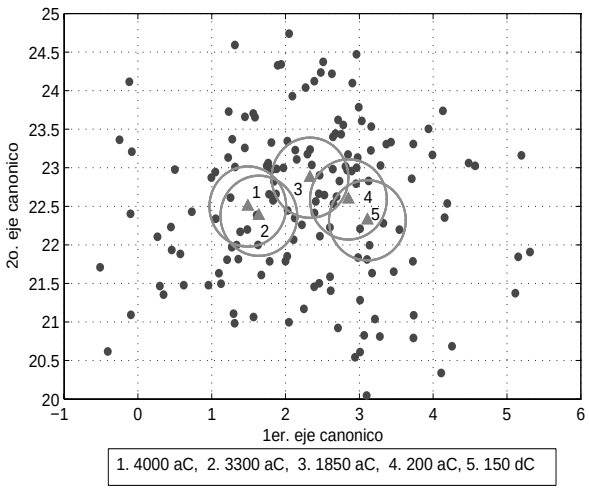


Figura 8.4.
Regiones confidenciales al 90% (Problema 8.3)

En las Figuras 8.3 y 8.4 puede observarse la ordenación temporal de los cinco períodos históricos a lo largo del primer eje canónico. Por tanto, este primer eje puede interpretarse como la evolución del cráneo a lo largo de la historia, con una tendencia hacia cráneos más anchos y algo achatados, con mandíbulas pequeñas y narices relativamente largas.

(d) La matriz de distancias entre los cinco grupos puede obtenerse a partir de las distancias euclídeas entre las filas de la matriz `mY`, que contiene las coordenadas de los individuos medios en función de las coordenadas canónicas:

```
squareform(pdist(mY))=      0    0.1920    0.9216    1.3660    1.6303
                       0.1920         0    0.8507    1.2317    1.4719
                       0.9216    0.8507         0    0.5913    0.9535
                       1.3660    1.2317    0.5913         0    0.3736
                       1.6303    1.4719    0.9535    0.3736         0
```

Podéis comparar los elementos de esta matriz con la representación canónica de los individuos medios que se muestra en la Figura 8.5. Recordad que las distancias representadas en la Figura 8.5 coinciden con las distancias de Mahalanobis entre los individuos medios en función de las variables originales. Por tanto, para estudiar posibles relaciones entre los distintos grupos será más cómodo observar el gráfico de la representación canónica que la matriz de distancias de Mahalanobis.

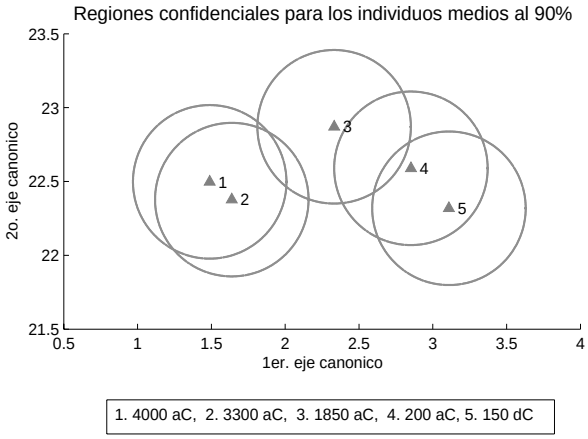


Figura 8.5.
Representación de las distancias entre los individuos medios (Problema 8.3)

PROBLEMA 8.4

*Las Tablas 8.3, 8.4 y 8.5 contienen varias variables medidas sobre 250 olmos, divididos en 3 grupos, según su sexo (*Grupo 1*: 100 olmos femeninos,* Grupo 2*: 100 olmos masculinos,* Grupo 3*: 50 olmos juveniles o plántulas). Véase el Problema 4.5 para una descripción completa de las variables.*

(a) *Realícese la representación canónica de los tres grupos, especificando los porcentajes de variabilidad explicados por cada eje canónico.*

(b) *Suponiendo normalidad multivariante, constrúyanse las regiones confidenciales (al 95%) para los individuos medios de cada grupo.*

(c) *Interprétense los ejes canónicos.*

Tabla 8.3.
Datos para el Problema 8.4. Grupo 1: olmos femeninos.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_1	X_2	X_3	X_4	X_5	X_6	X_7
0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	0.53	0.415	0.115	0.5915	0.233	0.1585	0.18
0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	0.49	0.375	0.135	0.6125	0.2555	0.102	0.22
0.545	0.425	0.125	0.768	0.294	0.1495	0.26	0.56	0.43	0.15	0.8825	0.3465	0.172	0.31
0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	0.47	0.365	0.105	0.4205	0.163	0.1035	0.14
0.525	0.38	0.14	0.6065	0.194	0.1475	0.21	0.515	0.425	0.14	0.766	0.304	0.1725	0.255
0.535	0.405	0.145	0.6845	0.2725	0.171	0.205	0.44	0.35	0.125	0.4035	0.175	0.063	0.129
0.47	0.355	0.1	0.4755	0.1675	0.0805	0.185	0.325	0.26	0.09	0.1915	0.085	0.036	0.062
0.44	0.34	0.1	0.451	0.188	0.087	0.13	0.425	0.33	0.115	0.406	0.1635	0.081	0.1355
0.565	0.44	0.155	0.9395	0.4275	0.214	0.27	0.305	0.23	0.08	0.156	0.0675	0.0345	0.048
0.55	0.415	0.135	0.7635	0.318	0.21	0.2	0.405	0.325	0.11	0.3555	0.151	0.063	0.117
0.615	0.48	0.165	1.1615	0.513	0.301	0.305	0.565	0.445	0.155	0.826	0.341	0.2055	0.2475
0.56	0.44	0.14	0.9285	0.3825	0.188	0.3	0.55	0.45	0.145	0.741	0.295	0.1435	0.2665
0.58	0.45	0.185	0.9955	0.3945	0.272	0.285	0.49	0.38	0.125	0.549	0.245	0.1075	0.174
0.68	0.56	0.165	1.639	0.6055	0.2805	0.46	0.605	0.5	0.185	1.1185	0.469	0.2585	0.335
0.68	0.55	0.175	1.798	0.815	0.3925	0.455	0.635	0.515	0.19	1.3715	0.5065	0.305	0.45
0.705	0.55	0.2	1.7095	0.633	0.4115	0.49	0.605	0.485	0.16	1.0565	0.37	0.2355	0.355
0.54	0.475	0.155	1.217	0.5305	0.3075	0.34	0.565	0.45	0.135	0.9885	0.387	0.1495	0.31
0.45	0.355	0.105	0.5225	0.237	0.1165	0.145	0.575	0.46	0.19	0.994	0.392	0.2425	0.34
0.575	0.445	0.135	0.883	0.381	0.2035	0.26	0.58	0.455	0.17	0.9075	0.374	0.2135	0.285
0.45	0.335	0.105	0.425	0.1865	0.091	0.115	0.575	0.46	0.165	1.124	0.2985	0.1785	0.44
0.55	0.425	0.135	0.8515	0.362	0.196	0.27	0.605	0.485	0.16	1.222	0.53	0.2575	0.28
0.46	0.375	0.12	0.4605	0.1775	0.11	0.15	0.725	0.56	0.21	2.141	0.65	0.398	1.005
0.525	0.425	0.16	0.8355	0.3545	0.2135	0.245	0.65	0.545	0.23	1.752	0.5605	0.2895	0.815
0.47	0.36	0.12	0.4775	0.2105	0.1055	0.15	0.725	0.575	0.175	2.124	0.765	0.4515	0.85
0.5	0.4	0.14	0.6615	0.2565	0.1755	0.22	0.68	0.57	0.205	1.842	0.625	0.408	0.65
0.505	0.4	0.125	0.583	0.246	0.13	0.175	0.68	0.515	0.175	1.6185	0.5125	0.409	0.62
0.53	0.41	0.13	0.6965	0.302	0.1935	0.2	0.53	0.395	0.145	0.775	0.308	0.169	0.255
0.565	0.44	0.16	0.915	0.354	0.1935	0.32	0.52	0.405	0.115	0.776	0.32	0.1845	0.22
0.595	0.495	0.185	1.285	0.416	0.224	0.485	0.56	0.45	0.16	1.0235	0.429	0.268	0.3
0.475	0.39	0.12	0.5305	0.2135	0.1155	0.17	0.62	0.475	0.175	1.0165	0.4355	0.214	0.325
0.4	0.32	0.11	0.353	0.1405	0.0985	0.1	0.645	0.51	0.2	1.5675	0.621	0.367	0.46
0.595	0.475	0.17	1.247	0.48	0.225	0.425	0.63	0.48	0.15	1.0525	0.392	0.336	0.285
0.605	0.45	0.195	1.098	0.481	0.2895	0.315	0.63	0.5	0.185	1.383	0.54	0.3315	0.38
0.6	0.475	0.15	1.0075	0.4425	0.221	0.28	0.63	0.48	0.16	1.199	0.5265	0.335	0.315
0.6	0.47	0.15	0.922	0.363	0.194	0.305	0.585	0.46	0.17	0.9325	0.365	0.271	0.29
0.555	0.425	0.14	0.788	0.282	0.1595	0.285	0.51	0.4	0.14	0.8145	0.459	0.1965	0.195
0.615	0.475	0.17	1.1025	0.4695	0.2355	0.345	0.505	0.41	0.15	0.644	0.285	0.145	0.21
0.575	0.445	0.14	0.941	0.3845	0.252	0.285	0.45	0.345	0.12	0.4165	0.1655	0.095	0.135
0.52	0.425	0.165	0.9885	0.396	0.225	0.32	0.5	0.4	0.145	0.63	0.234	0.1465	0.23
0.57	0.465	0.18	1.295	0.339	0.2225	0.44	0.53	0.435	0.17	0.8155	0.2985	0.155	0.275
0.46	0.355	0.13	0.517	0.2205	0.114	0.165	0.44	0.34	0.14	0.482	0.186	0.1085	0.16
0.575	0.45	0.16	0.9775	0.3135	0.231	0.33	0.525	0.415	0.17	0.8325	0.2755	0.1685	0.31
0.625	0.495	0.165	1.262	0.507	0.318	0.39	0.49	0.365	0.145	0.6345	0.1995	0.1625	0.22
0.475	0.375	0.125	0.5785	0.2775	0.085	0.155	0.415	0.325	0.105	0.38	0.1595	0.0785	0.12
0.52	0.41	0.155	0.727	0.291	0.1835	0.235	0.485	0.395	0.16	0.66	0.2475	0.128	0.235
0.545	0.43	0.165	0.802	0.2935	0.183	0.28	0.415	0.305	0.13	0.32	0.1305	0.0755	0.105
0.5	0.4	0.125	0.6675	0.261	0.1315	0.22	0.445	0.325	0.125	0.455	0.1785	0.1125	0.14
0.51	0.39	0.135	0.6335	0.231	0.179	0.2	0.47	0.35	0.145	0.5175	0.187	0.1235	0.18
0.435	0.395	0.105	0.3635	0.136	0.098	0.13	0.49	0.375	0.15	0.5755	0.22	0.144	0.19
0.545	0.41	0.125	0.6935	0.2975	0.146	0.21	0.445	0.355	0.15	0.485	0.181	0.125	0.155

SOLUCIÓN

(a) Sea `X=[X1;X2;X3]` la matriz 250×7 que contiene los datos de las Tablas 8.3, 8.4 y 8.5, en este orden. Para realizar el análisis canónico de poblaciones utilizaremos la función `canp.m`:

```
n = [100 100 50];
[mY,V,B,W,percent,Test1,texto1,Test2,texto2] = canp(X,n)
```

El vector `percent` contiene los porcentajes de variabilidad explicados por los 2 ejes ca-

Tabla 8.4.
Datos para el Problema 8.4. Grupo 2: olmos masculinos

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_1	X_2	X_3	X_4	X_5	X_6	X_7
0.665	0.525	0.165	1.338	0.5515	0.3575	0.35	0.515	0.405	0.13	0.722	0.32	0.131	0.21
0.465	0.355	0.105	0.4795	0.227	0.124	0.125	0.645	0.485	0.215	1.514	0.546	0.2615	0.635
0.355	0.29	0.09	0.3275	0.134	0.086	0.09	0.605	0.465	0.165	1.056	0.4215	0.2475	0.34
0.47	0.37	0.12	0.5795	0.293	0.227	0.14	0.61	0.485	0.175	1.2445	0.544	0.297	0.345
0.4	0.32	0.095	0.303	0.1335	0.06	0.1	0.725	0.57	0.19	2.55	1.0705	0.483	0.725
0.485	0.36	0.13	0.5415	0.2595	0.096	0.16	0.705	0.56	0.22	1.981	0.8175	0.3085	0.76
0.405	0.31	0.1	0.385	0.173	0.0915	0.11	0.695	0.55	0.215	1.9565	0.7125	0.541	0.59
0.445	0.35	0.12	0.4425	0.192	0.0955	0.135	0.525	0.435	0.155	1.065	0.486	0.233	0.285
0.47	0.385	0.135	0.5895	0.2765	0.12	0.17	0.58	0.475	0.15	0.97	0.385	0.2165	0.35
0.45	0.345	0.105	0.4115	0.18	0.1125	0.135	0.57	0.48	0.18	0.9395	0.399	0.2	0.295
0.505	0.405	0.11	0.625	0.305	0.16	0.175	0.64	0.51	0.175	1.368	0.515	0.266	0.57
0.425	0.325	0.095	0.3785	0.1705	0.08	0.1	0.62	0.49	0.19	1.218	0.5455	0.2965	0.355
0.52	0.4	0.12	0.58	0.234	0.1315	0.185	0.615	0.48	0.18	1.1595	0.4845	0.2165	0.325
0.475	0.355	0.12	0.48	0.234	0.1015	0.135	0.61	0.485	0.17	1.0225	0.419	0.2405	0.36
0.555	0.425	0.13	0.7665	0.264	0.168	0.275	0.58	0.45	0.15	0.927	0.276	0.1815	0.36
0.57	0.48	0.175	1.185	0.474	0.261	0.38	0.5	0.405	0.155	0.772	0.346	0.1535	0.245
0.595	0.475	0.14	0.944	0.3625	0.189	0.315	0.64	0.5	0.185	1.3035	0.4445	0.2635	0.465
0.62	0.51	0.175	1.615	0.5105	0.192	0.675	0.56	0.45	0.16	0.922	0.432	0.178	0.26
0.595	0.475	0.16	1.3175	0.408	0.234	0.58	0.585	0.46	0.185	0.922	0.3635	0.213	0.285
0.58	0.45	0.14	1.013	0.38	0.216	0.36	0.5	0.4	0.165	0.825	0.254	0.205	0.285
0.625	0.465	0.14	1.195	0.4825	0.205	0.4	0.42	0.335	0.115	0.369	0.171	0.071	0.12
0.56	0.44	0.16	0.8645	0.3305	0.2075	0.26	0.335	0.25	0.09	0.181	0.0755	0.0415	0.06
0.565	0.425	0.135	0.8115	0.341	0.1675	0.255	0.5	0.405	0.14	0.6155	0.241	0.1355	0.205
0.555	0.44	0.15	0.755	0.307	0.1525	0.26	0.55	0.405	0.14	0.8025	0.244	0.1635	0.255
0.595	0.465	0.175	1.115	0.4015	0.254	0.39	0.45	0.35	0.13	0.46	0.174	0.111	0.135
0.695	0.56	0.19	1.494	0.588	0.3425	0.485	0.47	0.36	0.135	0.501	0.1665	0.115	0.165
0.665	0.535	0.195	1.606	0.5755	0.388	0.48	0.555	0.445	0.135	0.836	0.336	0.1625	0.275
0.535	0.435	0.15	0.725	0.269	0.1385	0.25	0.565	0.44	0.175	0.9025	0.31	0.193	0.325
0.47	0.375	0.13	0.523	0.214	0.132	0.145	0.625	0.505	0.215	1.4455	0.496	0.287	0.435
0.47	0.37	0.13	0.5225	0.201	0.133	0.165	0.565	0.425	0.16	0.9425	0.3495	0.2185	0.275
0.55	0.435	0.145	0.843	0.328	0.1915	0.255	0.59	0.47	0.18	1.1235	0.4205	0.2805	0.36
0.53	0.435	0.16	0.883	0.316	0.164	0.335	0.6	0.495	0.165	1.2415	0.485	0.2775	0.34
0.53	0.415	0.14	0.724	0.3105	0.1675	0.205	0.56	0.45	0.175	1.011	0.3835	0.2065	0.37
0.605	0.47	0.16	1.1735	0.4975	0.2405	0.345	0.56	0.45	0.185	1.07	0.3805	0.175	0.41
0.495	0.395	0.125	0.5415	0.2375	0.1345	0.155	0.545	0.46	0.16	0.8975	0.341	0.1655	0.345
0.465	0.36	0.105	0.431	0.172	0.107	0.175	0.53	0.42	0.165	0.8945	0.319	0.239	0.245
0.425	0.35	0.105	0.393	0.13	0.063	0.165	0.27	0.2	0.08	0.1205	0.0465	0.028	0.04
0.44	0.34	0.105	0.402	0.1305	0.0955	0.165	0.52	0.45	0.15	0.895	0.3615	0.186	0.235
0.405	0.305	0.085	0.2605	0.1145	0.0595	0.085	0.35	0.275	0.11	0.2925	0.1225	0.0635	0.0905
0.37	0.265	0.075	0.214	0.09	0.051	0.07	0.47	0.39	0.15	0.6355	0.2185	0.0885	0.255
0.7	0.535	0.16	1.7255	0.63	0.2635	0.54	0.59	0.5	0.2	1.187	0.412	0.2705	0.37
0.71	0.54	0.165	1.959	0.7665	0.261	0.78	0.62	0.485	0.205	1.219	0.3875	0.2505	0.385
0.595	0.48	0.165	1.262	0.4835	0.283	0.41	0.63	0.505	0.225	1.525	0.56	0.3335	0.45
0.345	0.255	0.09	0.2005	0.094	0.0295	0.063	0.63	0.515	0.155	1.259	0.4105	0.197	0.41
0.375	0.285	0.095	0.253	0.096	0.0575	0.0925	0.655	0.54	0.215	1.844	0.7425	0.327	0.585
0.65	0.52	0.19	1.3445	0.519	0.306	0.4465	0.61	0.5	0.24	1.642	0.532	0.3345	0.69
0.56	0.455	0.155	0.797	0.34	0.19	0.2425	0.635	0.525	0.205	1.484	0.55	0.3115	0.43
0.475	0.375	0.13	0.5175	0.2075	0.1165	0.17	0.485	0.395	0.14	0.6295	0.2285	0.127	0.225
0.46	0.35	0.12	0.515	0.224	0.108	0.1565	0.515	0.38	0.175	0.9565	0.325	0.158	0.31
0.59	0.475	0.145	1.053	0.4415	0.262	0.325	0.53	0.435	0.155	0.699	0.288	0.1595	0.205

nónicos, con un 97.6454% de variabilidad explicada por el primer eje. Para el contraste de comparación de medias se obtiene una $F(14, 482) = 22.2795$, por lo que se infiere que las medias son distintas.

```
percent = 97.6454    2.3546
Test1 = 22.2795   14.0000  482.0000    0
texto1 = Test1: Igualdad de medias (Lambda de Wilks): p-valor= 0
```

La Figura 8.6 muestra la representación canónica de los individuos en dos dimensiones con un

Tabla 8.5.
Datos para el Problema 8.4. Grupo 3: olmos juveniles o plántulas.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_1	X_2	X_3	X_4	X_5	X_6	X_7
0.28	0.205	0.08	0.127	0.052	0.039	0.042	0.33	0.255	0.085	0.1655	0.063	0.039	0.06
0.175	0.13	0.055	0.0315	0.0105	0.0065	0.0125	0.35	0.26	0.085	0.174	0.0705	0.0345	0.06
0.17	0.13	0.095	0.03	0.013	0.008	0.01	0.32	0.245	0.08	0.1585	0.0635	0.0325	0.05
0.235	0.16	0.04	0.048	0.0185	0.018	0.015	0.36	0.275	0.085	0.1975	0.0745	0.0415	0.07
0.36	0.26	0.09	0.1785	0.0645	0.037	0.075	0.305	0.245	0.075	0.156	0.0675	0.038	0.045
0.315	0.21	0.06	0.125	0.06	0.0375	0.035	0.345	0.27	0.11	0.2135	0.082	0.0545	0.07
0.315	0.245	0.085	0.1435	0.053	0.0475	0.05	0.33	0.25	0.105	0.1715	0.0655	0.035	0.06
0.225	0.16	0.045	0.0465	0.025	0.015	0.015	0.245	0.195	0.06	0.095	0.0445	0.0245	0.026
0.355	0.275	0.085	0.22	0.092	0.06	0.15	0.36	0.285	0.105	0.2415	0.0915	0.057	0.075
0.4	0.3	0.11	0.315	0.109	0.067	0.12	0.295	0.215	0.085	0.128	0.049	0.034	0.04
0.435	0.34	0.11	0.3795	0.1495	0.085	0.12	0.275	0.205	0.075	0.1105	0.045	0.0285	0.035
0.37	0.28	0.095	0.2655	0.122	0.052	0.08	0.28	0.21	0.085	0.1065	0.039	0.0295	0.03
0.405	0.3	0.12	0.324	0.1265	0.07	0.11	0.2	0.145	0.06	0.037	0.0125	0.0095	0.011
0.425	0.38	0.105	0.3265	0.1285	0.0785	0.1	0.165	0.12	0.03	0.0215	0.007	0.005	0.005
0.365	0.27	0.085	0.205	0.078	0.0485	0.07	0.45	0.355	0.11	0.4585	0.194	0.067	0.14
0.275	0.215	0.075	0.1155	0.0485	0.029	0.035	0.33	0.255	0.095	0.172	0.066	0.0255	0.06
0.44	0.35	0.135	0.435	0.1815	0.083	0.125	0.265	0.21	0.06	0.0965	0.0425	0.022	0.03
0.295	0.225	0.08	0.124	0.0485	0.032	0.04	0.19	0.145	0.04	0.038	0.0165	0.0065	0.015
0.075	0.055	0.01	0.002	0.001	0.0005	0.0015	0.265	0.205	0.07	0.1055	0.039	0.041	0.035
0.13	0.1	0.03	0.013	0.0045	0.003	0.004	0.355	0.275	0.09	0.251	0.097	0.053	0.08
0.11	0.09	0.03	0.008	0.0025	0.002	0.003	0.32	0.255	0.1	0.1755	0.073	0.0415	0.065
0.16	0.12	0.035	0.021	0.0075	0.0045	0.005	0.36	0.28	0.09	0.2255	0.0885	0.04	0.09
0.27	0.2	0.07	0.1	0.034	0.0245	0.035	0.3	0.22	0.08	0.121	0.0475	0.042	0.035
0.23	0.175	0.065	0.0645	0.026	0.0105	0.02	0.235	0.175	0.04	0.0705	0.0335	0.015	0.02
0.3	0.23	0.08	0.1275	0.0435	0.0265	0.04	0.34	0.26	0.08	0.2	0.08	0.0555	0.055

100% de la variabilidad explicada.

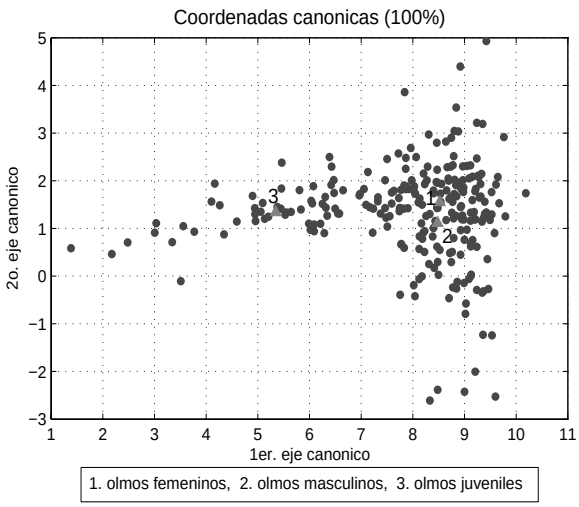


Figura 8.6.
Análisis canónico de poblaciones (Problema 8.4)

(b) Las regiones confidenciales para los individuos medios, al 95% son:

```
r = regconf(mY,n,7,0.95)
r =  0.3833    0.3833    0.5420
```

Puesto que el número de individuos es considerable, para una interpretación más clara, representaremos solamente los individuos medios y sus regiones confidenciales. La Figura 8.7 contiene esta representación, donde puede observarse que las diferencias entre los tres grupos son debidas al grupo de olmos juveniles.

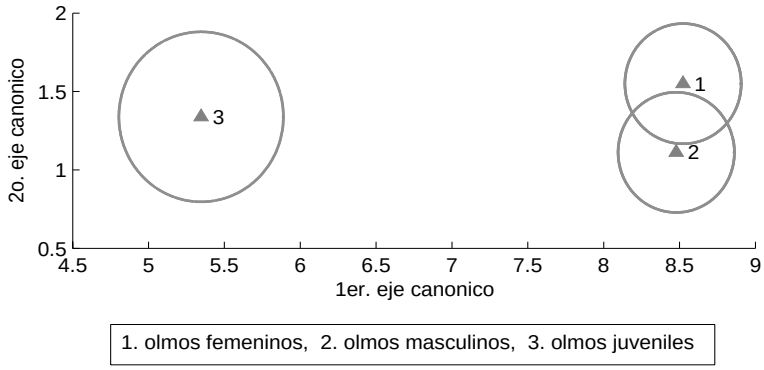


Figura 8.7.
Regiones confidenciales al 95% (Problema 8.4)

(c) Las columnas de la matriz V contienen los coeficientes de los ejes canónicos. Así el primer eje es:

$$Y_1 = 11.66\,X_1 + 8.38\,X_2 + 5.35\,X_3 - 3.20\,X_4 + 2.49\,X_5 - 1.31\,X_6 + 0.27\,X_7,$$

donde las variables que más contribuyen son X_1, X_2 y X_3, que corresponden, respectivamente, a la longitud (o mayor medida de la corteza), el diámetro y la altura (con madera dentro de la corteza) del olmo. Contribuciones positivas en estas tres variables indican un árbol alto y grueso, por tanto, el primer eje canónico podría interpretarse como una medida general del tamaño del olmo. En la Figura 8.7 se observa que los olmos juveniles son los que puntúan más bajo respecto del primer eje canónico, mientras que los olmos adultos (femeninos o masculinos) puntúan más alto. También se observa que entre los olmos adultos no existen diferencias en cuanto a su tamaño, pero sí respecto del segundo eje canónico:

$$Y_2 = 18.11\,X_1 - 10.74\,X_2 - 18.86\,X_3 - 8.04\,X_4 - 4.14\,X_5 + 28.47\,X_6 + 7.11\,X_7.$$

Las variables que más contribuyen a este eje son X_1, X_2, X_3 y X_6, que es el peso de las vísceras. La primera y última variables lo hacen en sentido positivo, mientras que X_2 y X_3 lo hacen en sentido negativo. Este segundo eje podría interpretarse como un índice del contenido relativo de madera del árbol. En la Figura 8.7 se observa que los olmos femeninos puntúan más alto respecto de este eje, indicando que el contenido relativo de madera es mayor en este grupo de individuos.

PROBLEMA 8.5

La Tabla 8.6 contiene once variables medidas sobre un total de 44 individuos pertenecientes a cuatro especies de cocodrilos: 1. Alligator mississippiensis*, 2.* Crocodylus niloticus*, 3.* Crocodylus porosus*, 4.* Osteolaemus tetraspis*. La Figura 8.8 muestra las regiones geográficas donde se encuentran estas especies de cocodrilos. Las variables medidas sobre cada individuo son: X_1 =longitud del cráneo, X_2 =ancho del cráneo, X_3 =ancho del hocico, X_4 =longitud del hocico, X_5 =longitud dorsal del cráneo, X_6 =ancho máximo orbital, X_7 =ancho mínimo inter-orbital, X_8 =longitud máxima orbital, X_9 =longitud del paladar post-orbital, X_{10} =ancho posterior del paladar, X_{11} =ancho máximo entre orificios nasales (Fuente: Iordansky 1973).*

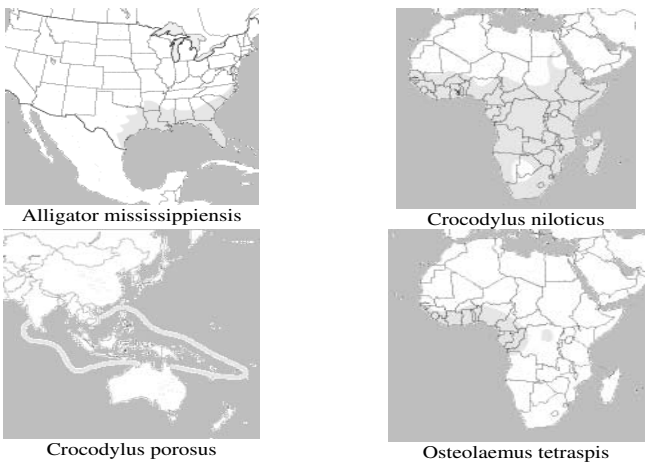


Figura 8.8.
Hábitat de las cuatro especies de cocodrilos. (Problema 8.5)

Realícese la representación canónica de las cuatro especies, especificando los porcentajes de variabilidad explicados por cada eje canónico. Suponiendo normalidad multivariante, constrúyanse las regiones confidenciales (al 90%) para los individuos medios de cada grupo.

SOLUCIÓN

Sea `X` la matriz que contiene los datos de la Tabla 8.6. Para poder utilizar las funciones `canp` y `regconf`, construimos mediante la función interna de Matlab `find`, un vector que contenga el número de individuos de cada grupo:

```
n = zeros(1,4);
for i = 1:4
   grupo = find(X(:,1)==i);
   n(i) = length(grupo);
end
```

Tabla 8.6.
Datos para el Problema 8.5.

especie	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1	72	40	37	35	71	17	5	20	15	25	11
1	220	112	98	138	216	30	16	46	36	64	31
1	225	150	89	140	220	32	17	52	37	82	30
1	272	138	120	175	262	24	25	54	44	78	38
1	288	148	126	180	275	40	22	58	42	82	40
1	290	150	117	183	270	40	20	54	46	82	40
1	292	150	127	166	284	49	26	56	48	86	39
1	320	150	124	203	310	40	25	62	46	80	38
1	354	178	137	240	337	42	25	69	50	89	51
1	366	186	160	232	348	39	32	68	54	98	53
1	380	236	210	238	358	52	27	63	63	120	64
2	160	64	46	100	153	20	9	22	30	39	9
2	198	94	70	121	186	25	13	31	32	48	13
2	248	243	76	159	235	30	16	41	42	105	15
2	254	114	71	158	235	28	16	40	42	65	15
2	420	235	170	270	400	37	42	60	68	105	42
2	440	250	170	280	420	42	50	65	70	120	48
2	525	290	220	360	495	45	48	72	82	145	54
2	582	336	218	382	554	48	58	72	76	105	57
2	610	345	268	400	564	46	90	85	76	164	56
3	76	30	22	41	73	13	4	17	16	20	4
3	548	74	56	364	513	23	10	29	26	44	48
3	238	292	68	154	230	29	12	36	30	55	48
3	408	200	148	274	390	38	36	57	54	110	32
3	548	300	210	364	513	46	55	68	65	150	48
3	565	292	216	405	550	45	64	70	90	160	48
3	672	384	302	452	620	50	70	90	85	185	64
3	800	416	324	516	740	63	82	100	105	204	75
4	164	90	70	90	160	36	16	42	32	57	20
4	188	107	71	92	160	29	13	38	35	65	18
4	170	98	72	98	165	31	14	42	35	60	20
4	173	107	70	100	165	33	12	40	35	60	22
4	175	102	73	102	165	32	14	42	38	64	24
4	185	105	77	105	175	32	14	44	40	61	22
4	185	105	78	105	175	33	16	40	40	61	22
4	188	107	82	108	180	33	16	40	40	65	24
4	188	104	80	110	178	34	15	44	40	64	24
4	190	108	80	112	180	32	16	45	38	65	24
4	194	110	82	114	182	34	15	44	38	67	24
4	194	117	92	117	180	34	18	43	42	70	23
4	203	108	88	116	193	35	16	46	40	69	26
4	210	107	91	124	178	36	19	48	40	65	26
4	225	128	105	128	215	40	20	52	45	75	28
4	240	136	91	133	222	38	19	51	46	76	27

y obtenemos: `n=[11 9 8 16]`. Para obtener la representación canónica, haremos

```
[mY,V,B,W,percent,Test1,texto1,Test2,texto2] = canp(X(:,2:11),n)
r = regconf(mY,n,11,0.90)
```

El vector `percent` contiene los porcentajes de variabilidad explicados por los 2 ejes canónicos, con un 77.0181% de variabilidad explicada por el primer eje. Para el contraste de compa-

ración de medias se obtiene una $F(33, 89) = 6.2528$, por lo que se infiere que las medias son distintas.

```
mY =
 5.2342    0.3893
-0.5270   -2.1787
-1.0138   -0.1066
 4.2779   -2.7562

percent = 77.0181   20.4787    2.5032
Test1 = 6.2528   33.0000   89.0000    0.0000
texto1 = Test1: Igualdad de medias (Lambda de Wilks): p-valor=2.6755e-012
```

La Figura 8.9 contiene la representación canónica de los cocodrilos en dos dimensiones con un 97.5% de la variabilidad explicada junto con las regiones confidenciales al 90% para los individuos medios.

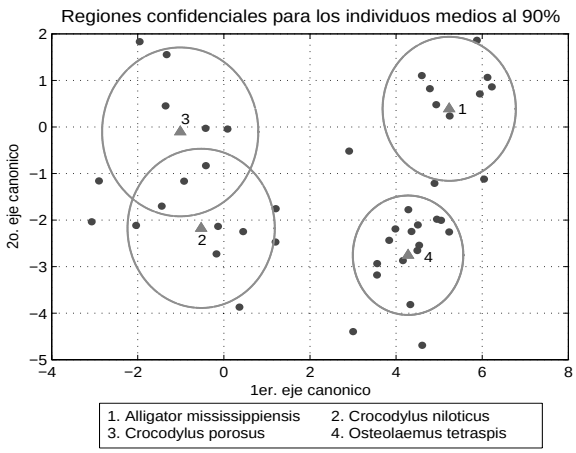


Figura 8.9.
Análisis canónico de poblaciones (Problema 8.5)

CAPÍTULO 9

Análisis discriminante y clasificación

Supongamos que tenemos varias poblaciones conocidas $\Omega_1, \ldots, \Omega_g$, en cada una de las cuales observamos una muestra de cierto vector de interés $\mathbf{X} = (X_1, \ldots, X_p)'$.
El análisis discriminante se ocupa de describir, mediante las variables X_i, los rasgos diferenciales entre las poblaciones. Se trata de encontrar *funciones discriminantes* o *reglas de decisión* $h = h(x_1, \ldots, x_p)$ cuyos valores en los distintos grupos estén lo más separados posible. O, más precisamente, buscamos funciones h sencillas que permitan asignar cada observación $\mathbf{x} = (x_1, \ldots, x_p)'$ a una población Ω_i minimizando la tasa de error en dicha asignación. La más conocida es la regla discriminante lineal de Fisher, donde h es una función lineal de $\mathbf{x}$.
El problema de clasificación, como su mismo nombre indica, trata de clasificar una nueva observación $\mathbf{x}$, cuya población de procedencia se desconoce, en alguna de las poblaciones Ω_i. Para ello se utilizan las funciones discriminantes construidas a partir de la muestra.

PROBLEMA 9.1

Sean Ω_1 y Ω_2 dos poblaciones con distribuciones $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ y $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ respectivamente. El discriminador lineal de Fisher que asigna $\mathbf{x} \in \mathbb{R}^p$ a una de las dos poblaciones anteriores es

$$L(\mathbf{x}) = \left(\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

La regla discriminante lineal de Fisher consiste en asignar $\mathbf{x}$ a la población Ω_1 si $L(\mathbf{x}) > 0$ y, en caso contrario, asignar $\mathbf{x}$ a la población Ω_2.

(a) *Exprésese $L(\mathbf{x})$ como la diferencia entre los cuadrados de las distancias de Mahalanobis de $\mathbf{x}$ a $\boldsymbol{\mu}_1$ y de $\mathbf{x}$ a $\boldsymbol{\mu}_2$.*

(b) *Demuéstrese que la probabilidad de clasificación errónea es $pce = \Phi(-M/2)$, donde $M^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ es la distancia de Mahalanobis entre las poblaciones Ω_1 y Ω_2 y Φ es la función de distribución de una ley normal estándar.*

SOLUCIÓN

(a)
$$\begin{aligned} L(\mathbf{x}) &= \mathbf{x}'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_1 - \mathbf{x}'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_2 - \frac{1}{2}\,\boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_1 \\ &+ \frac{1}{2}\,\boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_2 - \frac{1}{2}\,\boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_1 + \frac{1}{2}\,\boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_2 \\ &= \frac{1}{2}\,\mathbf{x}'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_1 + \frac{1}{2}\,\mathbf{x}'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_1 - \frac{1}{2}\,\mathbf{x}'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_2 \\ &- \frac{1}{2}\,\mathbf{x}'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_2 - \frac{1}{2}\,\boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_1 + \frac{1}{2}\,\boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_2 \\ &= \frac{1}{2}\left(\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\right. \\ &- \left.\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\right) \\ &= \frac{1}{2}\left(d^2_{Mah}(\mathbf{x}, \mu_2) - d^2_{Mah}(\mathbf{x}, \boldsymbol{\mu}_1)\right). \end{aligned}$$

(b) El discriminador lineal de Fisher puede escribirse como

$$L(\mathbf{x}) = \left(\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)' \boldsymbol{\Sigma}^{-1}\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = (\mathbf{x} - \boldsymbol{\mu})'\,\mathbf{a} = \mathbf{a}'\,(\mathbf{x} - \boldsymbol{\mu}),$$

donde $\mathbf{a} = \boldsymbol{\Sigma}^{-1}\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ y $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Si $\mathbf{x} \in \mathbb{R}^p$ proviene de alguna de las poblaciones Ω_i, $i = 1, 2$, $L(\mathbf{x})$ tendrá ley normal. Su varianza será

$$var(L(\mathbf{x})) = var\,(\mathbf{a}'\,(\mathbf{x} - \boldsymbol{\mu})) = \mathbf{a}'\,\boldsymbol{\Sigma}\,\mathbf{a} = M^2$$

y su esperanza:

$$E(L(\mathbf{x})) = \mathbf{a}'\, E(\mathbf{x}-\boldsymbol{\mu}) = \begin{cases} \frac{1}{2}\,\mathbf{a}'\,(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = \frac{1}{2}\,M^2, & \text{si } \mathbf{x}\in\Omega_1, \\ -\frac{1}{2}\,\mathbf{a}'\,(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = -\frac{1}{2}\,M^2, & \text{si } \mathbf{x}\in\Omega_2. \end{cases}$$

Por tanto, $L(\mathbf{x}) \sim N(\frac{1}{2}M^2, M^2)$ si $\mathbf{x} \in \Omega_1$ y $L(\mathbf{x}) \sim N(-\frac{1}{2}M^2, M^2)$ si $\mathbf{x} \in \Omega_2$. El individuo $\mathbf{x}$ se clasificará erróneamente cuando se asigne a la población Ω_1 y en realidad provenga de Ω_2, o bien, cuando se asigne a la población Ω_2 y en realidad provenga de Ω_1. Luego la probabilidad de clasificación errónea es:

$$\begin{aligned} pce &= \frac{1}{2}\,P\left(L(\mathbf{x})>0/\mathbf{x}\in\Omega_2\right) + \frac{1}{2}\,P\left(L(\mathbf{x})<0/\mathbf{x}\in\Omega_1\right) \\ &= \frac{1}{2}\,P\left(\frac{L(\mathbf{x})+\frac{1}{2}M^2}{M} > \frac{\frac{1}{2}M^2}{M}\right) + \frac{1}{2}\,P\left(\frac{L(\mathbf{x})-\frac{1}{2}M^2}{M} < \frac{-\frac{1}{2}M^2}{M}\right) \\ &= \frac{1}{2}\,\Phi\left(-\frac{M}{2}\right) + \frac{1}{2}\,\Phi\left(-\frac{M}{2}\right) = \Phi\left(-\frac{M}{2}\right). \end{aligned}$$

PROBLEMA 9.2

Sean Ω_1 y Ω_2 dos poblaciones y $\mathbf{X} = (X_1, \ldots, X_p)'$ un vector con distribución de probabilidad conocida, dependiente de un parámetro $\boldsymbol{\theta}$ que toma el valor $\boldsymbol{\theta}_1$ si $\mathbf{X} \in \Omega_1$ y $\boldsymbol{\theta}_2$ si $\mathbf{X} \in \Omega_2$. Sea $\mathbf{x} = (x_1, \ldots, x_p)'$ el vector de observaciones de $\mathbf{X}$ sobre un individuo ω. La probabilidad o verosimilitud de la observación $\mathbf{x}$ en Ω_i es

$$\mathcal{L}_i(\mathbf{x}) = f(x_1, \ldots, x_p; \boldsymbol{\theta}_i).$$

La regla discriminante de máxima verosimilitud consiste en asignar ω a la población Ω_i para la que la verosimilitud de la observación es mayor. Esta regla tiene asociada la siguiente función discriminante

$$V(\mathbf{x}) = \log \mathcal{L}_1(\mathbf{x}) - \log \mathcal{L}_2(\mathbf{x}).$$

Dada una probabilidad a priori, $q_i = P(\omega \in \Omega_i)$, *la probabilidad* a posteriori, *conocido $\mathbf{x}$, se obtiene de la expresión:*

$$P(\omega \in \Omega_i|\mathbf{x}) = \frac{q_i\,\mathcal{L}_i(\mathbf{x})}{q_1\,\mathcal{L}_1(\mathbf{x}) + q_2\,\mathcal{L}_2(\mathbf{x})}.$$

La regla discriminante de Bayes asigna ω a la población Ω_i para la que $P(\omega \in \Omega_i|\mathbf{x})$ es máxima. Esta regla tiene asociada la siguiente función discriminante

$$B(\mathbf{x}) = \log \mathcal{L}_1(\mathbf{x}) - \log \mathcal{L}_2(\mathbf{x}) + \log(q_1/q_2).$$

Para este problema supóngase que Ω_i es una población $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$.

(a) *Demuéstrese que si* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ *la regla de máxima verosimilitud y la regla de Bayes con* $q_1 = q_2 = 1/2$ *coinciden con la regla discriminante lineal de Fisher (véase el Problema 9.1).*

(b) *Demuéstrese que si* $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$*, la regla de máxima verosimilitud está basada en el discriminador cuadrático*

$$\begin{aligned} Q(\mathbf{x}) &= \frac{1}{2}\mathbf{x}'\left(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}\right)\mathbf{x} + \mathbf{x}'\left(\boldsymbol{\Sigma}_1^{-1}\,\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\,\boldsymbol{\mu}_2\right) \\ &+ \frac{1}{2}\boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}_2^{-1}\,\boldsymbol{\mu}_2 - \frac{1}{2}\boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}_1^{-1}\,\boldsymbol{\mu}_1 + \frac{1}{2}\log|\boldsymbol{\Sigma}_2| - \frac{1}{2}\log|\boldsymbol{\Sigma}_1|. \end{aligned}$$

SOLUCIÓN

La función de densidad de $\mathbf{x}$ en la población Ω_i es:

$$f_i(\mathbf{x}) = \frac{|\boldsymbol{\Sigma}_i|^{-1/2}}{(2\,\pi)^{p/2}} \exp\left\{-\frac{1}{2}\,(\mathbf{x} - \boldsymbol{\mu}_i)'\,\boldsymbol{\Sigma}_i^{-1}\,(\mathbf{x} - \boldsymbol{\mu}_i)\right\}$$

(a) Supongamos que $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. La regla de máxima verosimilitud es

$$\begin{aligned} V(\mathbf{x}) &= \log f_1(\mathbf{x}) - \log f_2(\mathbf{x}) \\ &= -\frac{1}{2}\,\log\frac{|\boldsymbol{\Sigma}|}{(2\,\pi)^p} - \frac{1}{2}\,(\mathbf{x} - \boldsymbol{\mu}_1)'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x} - \boldsymbol{\mu}_1) \\ &+ \frac{1}{2}\,\log\frac{|\boldsymbol{\Sigma}|}{(2\,\pi)^p} + \frac{1}{2}\,(\mathbf{x} - \boldsymbol{\mu}_2)'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x} - \boldsymbol{\mu}_2) \\ &= \frac{1}{2}\Big((\mathbf{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) - (\mathbf{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\Big), \end{aligned}$$

que, como se demostró en el Problema 9.1, es una de las expresiones del discrimador lineal de Fisher.
La regla de Bayes con $q_1 = q_2 = 1/2$ (que implica $\log(q_1/q_2) = 0$) es

$$B(\mathbf{x}) = \log f_1(\mathbf{x}) - \log f_2(\mathbf{x}) + \log(q_1/q_2) = V(\mathbf{x}).$$

(b) Supongamos ahora que $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. La regla de máxima verosimilitud es

$$\begin{aligned} V(\mathbf{x}) &= \log f_1(\mathbf{x}) - \log f_2(\mathbf{x}) \\ &= -\frac{1}{2}\,\log|\boldsymbol{\Sigma}_1| - \frac{p}{2}\,\log(2\,\pi) - \frac{1}{2}\,(\mathbf{x} - \boldsymbol{\mu}_1)'\,\boldsymbol{\Sigma}_1^{-1}\,(\mathbf{x} - \boldsymbol{\mu}_1) \\ &+ \frac{1}{2}\,\log|\boldsymbol{\Sigma}_2| + \frac{p}{2}\,\log(2\,\pi) + \frac{1}{2}\,(\mathbf{x} - \boldsymbol{\mu}_2)'\,\boldsymbol{\Sigma}_2^{-1}\,(\mathbf{x} - \boldsymbol{\mu}_2) \\ &= -\frac{1}{2}\,\mathbf{x}'\,\boldsymbol{\Sigma}_1^{-1}\,\mathbf{x} + \mathbf{x}'\,\boldsymbol{\Sigma}_1^{-1}\,\boldsymbol{\mu}_1 - \frac{1}{2}\,\boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}_1^{-1}\,\boldsymbol{\mu}_1 + \frac{1}{2}\,\mathbf{x}'\,\boldsymbol{\Sigma}_2^{-1}\,\mathbf{x} \\ &- \mathbf{x}'\,\boldsymbol{\Sigma}_2^{-1}\,\boldsymbol{\mu}_2 + \frac{1}{2}\,\boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}_2^{-1}\,\boldsymbol{\mu}_2 + \frac{1}{2}\,\log|\boldsymbol{\Sigma}_2| - \frac{1}{2}\,\log|\boldsymbol{\Sigma}_1| \\ &= Q(\mathbf{x}). \end{aligned}$$

PROBLEMA 9.3

Se ha tomado una muestra de $n_1 = 25$ escuelas de formación artística y $n_2 = 25$ centros de investigación universitarios. En cada uno de ellos se ha observado un vector aleatorio $\mathbf{X} = (X_1, X_2)'$, donde X_1 es el porcentaje de profesores con grado de doctor en esa escuela o universidad. La variable X_2 es el porcentaje de estudiantes matriculados en el centro que estaban entre los mejores de su centro de educación secundaria. La muestra aparece representada en la Figura 9.1.

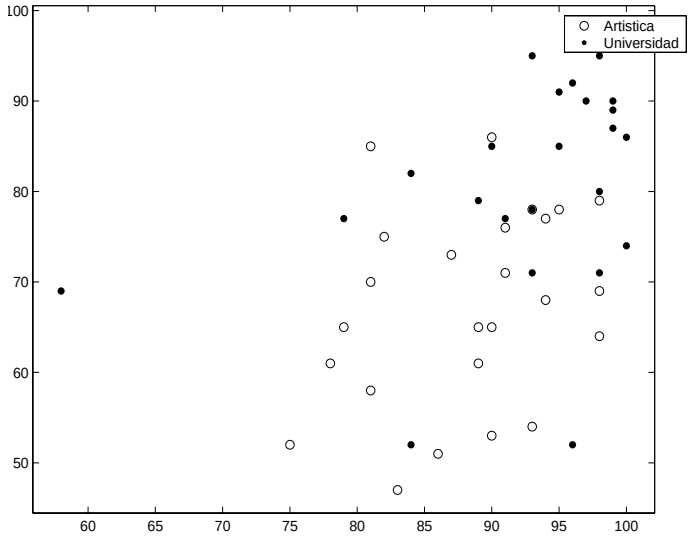


Figura 9.1.
Diagrama de dispersión con los datos del Problema 9.3

(a) *Dibújese sobre el gráfico la recta de proyección que en la opinión del lector (aproximadamente) mejor discrimine entre ambos grupos. Supóngase que la dirección de esa recta viene determinada por un vector $\mathbf{a} = (a_1, a_2)'$. Tratar de dar, a partir del dibujo, unos valores aproximados para a_1 y a_2.*

(b) *Ahora se quiere clasificar una nueva observación $\mathbf{x} = (x_1, x_2)'$ en alguno de los dos grupos: escuela de arte o centro de investigación. Sabiendo que los vectores de medias y matrices de covarianzas muestrales de ambas poblaciones son, respectivamente:*

$$\bar{\mathbf{x}}_1 = (88.24, 67.24)', \qquad \bar{\mathbf{x}}_2 = (92.88, 81.64)',$$

$$\mathbf{S}_1 = \begin{pmatrix} 44.35 & 22.73 \\ 22.73 & 116.69 \end{pmatrix}, \qquad \mathbf{S}_2 = \begin{pmatrix} 83.69 & 44.70 \\ 44.70 & 148.24 \end{pmatrix},$$

escríbase la regla de clasificación lineal de Fisher. Utilícese esta regla para asignar la observación $\mathbf{x} = (80, 60)'$ a una escuela de arte o a una universidad.

SOLUCIÓN

(a) Véase Figura 9.2.

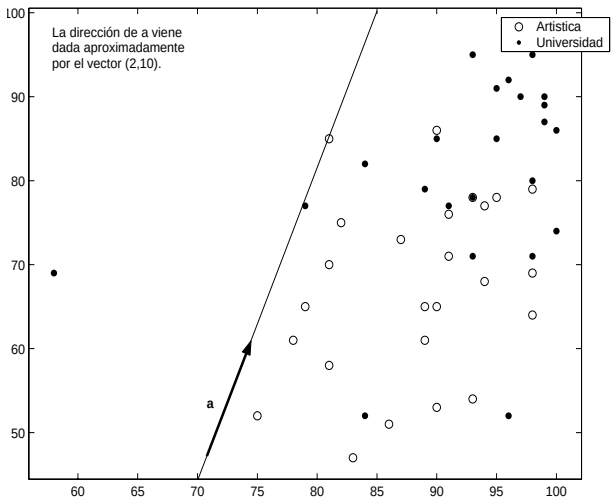


Figura 9.2.
Diagrama de dispersión y regla discriminante lineal (Problema 9.3)

(b) En este capítulo denotaremos por $\mathbf{S}_i$ la matriz de dispersión de la población i definida por

$$\mathbf{S}_i = \mathbf{X}_i'\mathbf{H}\mathbf{X}_i/(n_i - 1),$$

donde $\mathbf{X}_i$ es la matriz de datos de la población i y $\mathbf{H}$ es la matriz de centrado definida en el Problema 2.1. La regla discriminante lineal de Fisher asigna $\mathbf{x}$ a la población 1 (escuela de arte) si $\mathbf{a}'\mathbf{x} > m$, donde

$$\mathbf{a} = \mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = (-0.0176, -0.1042)',$$

$$\mathbf{S}_p = \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2 = \begin{pmatrix} 64.02 & 33.71 \\ 33.71 & 132.46 \end{pmatrix}$$

es la matriz de covarianzas común y

$$m = \frac{1}{2}\mathbf{a}'(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) = -9.35.$$

Puesto que $\mathbf{a}'\mathbf{x} = -7.66 > m$ asignamos $\mathbf{x} = (80, 60)'$ a la escuela de arte.

PROBLEMA 9.4

Un enólogo analiza dos componentes X_1 y X_2 en sendas muestras de dos tipos de vinos. Los resultados del análisis se pueden ver en la Tabla 9.1. Los datos se han extraído de Newman et al. (1998).

Tabla 9.1.
Muestras de dos vinos (Problema 9.4). Fuente: Newman *et al.* (1998)

Vino 1		Vino 2	
X_1	X_2	x_1	x_2
14.23	1065	12.37	520
13.20	1050	12.33	680
13.16	1185	12.64	450
14.37	1480	13.67	630
13.24	735	12.37	420
14.20	1450	12.17	355
14.39	1290	12.37	678
14.06	1295	13.11	502
14.83	1045	12.37	510
13.86	1045	13.34	750
14.10	1510	12.21	718
14.12	1280	12.29	870
13.75	1320	13.86	410
14.75	1150	13.49	472
14.38	1547	12.99	985
13.63	1310		
14.30	1280		
13.83	1130		
14.19	1680		
13.64	845		

(a) *Denotemos $\mathbf{X} = (X_1, X_2)'$. Exprésese la regla de clasificación lineal de Fisher para una nueva observación $\mathbf{x} = (x_1, x_2)'$. Prográmese como una función de Matlab.*

(b) *Aplíquese la regla de clasificación obtenida en el apartado anterior al caso concreto en que $\mathbf{x} = (13.05, 515)'$. ¿A qué tipo de vino corresponde?*

SOLUCIÓN

(a) La regla de Fisher está expresada en el apartado (b) del Problema 9.3. Para programarla como función de Matlab utilizaremos el siguiente código:

```
function poblacion = LinealDiscrim(x,X1,X2)

% LinealDiscrim(x,X1,X2)
% Clasifica el individuo x en la poblacion 1 o en la 2
% utilizando la regla discriminante lineal de Fisher.
% X1 y X2 son muestras de las poblaciones 1 y 2 respectivamente.

x = x(:) ; px = length(x) ;
```

```
[n1,p1] = size(X1) ; [n2,p2] = size(X2) ;
if p1 ~= p2
   error('Las matrices de datos no tienen dimensiones coherentes')
else
   p = p1 ;
   clear p1 p2
end
if px ~= p
   error('El vector x no tiene dimension adecuada')
else
   clear px
end

m1 = mean(X1) ; % Media muestral de poblacion 1
m2 = mean(X2) ; % Media muestral de poblacion 2
S1 = cov(X1) ; % Matriz de covarianzas (insesgado) de X1
S2 = cov(X2) ; % Matriz de covarianzas (insesgado) de X2
S_p = ((n1-1) * S1 + (n2-1) * S2)/(n1+n2-2); % Matriz de
                                              % covarianzas comun
a = S_p\((m1-m2)');
m = (m1+m2) * a/2;

if  (a'*x > m)
   poblacion = 1;
else
   poblacion = 2;
end
```

(b) Dado que `poblacion = LinealDiscrim(x,X1,X2)` toma el valor 2, asignaremos esta observación al segundo tipo de vino (que es de hecho la población de la que provenía).

PROBLEMA 9.5

La tabla 8.1 contiene cuatro medidas $\mathbf{X} = (X_1, X_2, X_3, X_4)'$ *sobre tres especies de flores del género* Iris *(véase el Problema 8.2 para una descripción completa de los datos). Supondremos que el vector* $\mathbf{X}$ *observado sigue una distribución normal. Dadas las tres nuevas flores (individuos)*

ind.	X_1	X_2	X_3	X_4
$\mathbf{x}_1$	*4.6*	*3.6*	*1.0*	*0.2*
$\mathbf{x}_2$	*6.8*	*2.8*	*4.8*	*1.4*
$\mathbf{x}_3$	*7.2*	*3.2*	*6.0*	*1.8*

*asígnense a alguna de las tres especies (*I. setosa, I. virginica *o* I. versicolor*) mediante*

(a) *el discriminador lineal,*

(b) *el discriminador cuadrático.*

SOLUCIÓN

Para asignar estos nuevos individuos a alguna de las poblaciones (especies) anteriores necesitamos una regla de decisión, que será distinta según el discriminador que se utilice. Para cualquiera de los dos métodos especificados en el enunciado debemos calcular los vectores de medias y las matrices de covarianzas de cada población y también el vector de medias global y la matriz de covarianzas común. Los vectores de medias son:

i	Población	$\bar{\mathbf{x}}_i'$			
1	*I. setosa*	5.01	3.43	1.46	0.25
2	*I. versicolor*	5.94	2.77	4.26	1.33
3	*I. virginica*	6.59	2.97	5.55	2.03

y las matrices de covarianzas de cada una de las especies son:

$\mathbf{S}_1$				$\mathbf{S}_2$				$\mathbf{S}_3$			
0.12	0.10	0.02	0.01	0.27	0.09	0.18	0.06	0.40	0.09	0.30	0.05
	0.14	0.01	0.01		0.10	0.08	0.04		0.10	0.07	0.05
		0.03	0.01			0.22	0.07			0.30	0.05
			0.01				0.04				0.08

Por tanto, la matriz de covarianzas común es:

$$\mathbf{S} = \begin{pmatrix} 0.27 & 0.10 & 0.17 & 0.04 \\ & 0.12 & 0.06 & 0.03 \\ & & 0.19 & 0.04 \\ & & & 0.04 \end{pmatrix}.$$

(a) La regla discriminante lineal asigna una nueva observación $\mathbf{x}$ a aquella población i tal que la distancia de Mahalanobis de $\mathbf{x}$ a su media $\bar{\mathbf{x}}_i$ sea mínima. La hipótesis que subyace es que la distribución de $\mathbf{X}$ es normal y tiene la misma matriz de covarianzas en todas las poblaciones. Por tanto, calcularemos

$$d(\mathbf{x}, \bar{\mathbf{x}}_i) = (\mathbf{x} - \bar{\mathbf{x}}_i)' \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i)$$

y asignaremos $\mathbf{x}$ a aquella población i tal que

$$d(\mathbf{x}, \bar{\mathbf{x}}_i) < d(\mathbf{x}, \bar{\mathbf{x}}_j)$$

para todo $i \neq j$.

De ahora en adelante suponemos que ya hemos definido en Matlab las matrices de datos `X1`, `X2` y `X3`, de dimensión 50×4 cada una, que contienen las observaciones de las especies *I. setosa*, *I. virginica* e *I. versicolor*, respectivamente.

La siguiente función permite realizar estos cálculos:

```
function [poblacion,D] = LinealDiscrim3(x,X1,X2,X3)

% [poblacion,D] = LinealDiscrim3(x,X1,X2,X3)
% Regla discriminante lineal para tres poblaciones
% Clasifica el individuo x en la poblacion 1, en la 2 o en la 3,
% utilizando la regla discriminante lineal.
% X1, X2 y X3 son muestras de las poblaciones 1, 2 y 3,
% respectivamente.
% D es el vector de distancias del individuo x a las
% poblaciones 1, 2 y 3.

x = x(:) ; px = length(x) ;
[n1,p1] = size(X1) ; [n2,p2] = size(X2) ; [n3,p3] = size(X3) ;
aux1 = [p1-p2, p2-p3, p1-p3] ;
if any(aux1 ~= 0)
    error('Las matrices de datos no tienen dimensiones coherentes')
else
    p = p1 ; clear p1 p2 p3
end
if px ~= p
    error('El vector x no tiene dimension adecuada')
else
    clear px
end

m1 = mean(X1) ; m2 = mean(X2) ; m3 = mean(X3) ;
S1 = cov(X1) ; S2 = cov(X2) ; S3 = cov(X3) ;
S = ((n1-1) * S1 + (n2-1) * S2 + (n3-1) * S3)/(n1+n2+n3-3) ;

x_rep = (ones(3,1) * x') - [ m1 ; m2 ; m3 ] ;
D = diag( x_rep * inv(S) * x_rep' ) ;
[Dmin,poblacion] = min(D) ;
```

La tabla siguiente muestra los vectores `D` para las tres nuevas flores:

ind.	$d(\mathbf{x}, \bar{\mathbf{x}}_1)$	$d(\mathbf{x}, \bar{\mathbf{x}}_2)$	$d(\mathbf{x}, \bar{\mathbf{x}}_3)$
$\mathbf{x}_1$	2.2864	113.6509	210.0239
$\mathbf{x}_2$	105.9403	3.7242	16.4216
$\mathbf{x}_3$	171.0985	17.3642	5.5252

luego clasificamos $\mathbf{x}_1$, $\mathbf{x}_2$ y $\mathbf{x}_3$ en las especies de Iris 1, 2 y 3, respectivamente.

(b) En el Problema 8.2 se vio que existían diferencias significativas entre las matrices de covarianzas. Así pues, el discriminador cuadrático podría resultar más adecuado en este caso. Esta regla de discriminación asigna la nueva observación $\mathbf{x}$ a la especie i si

$$d(\mathbf{x}, \bar{\mathbf{x}}_i) < d(\mathbf{x}, \bar{\mathbf{x}}_j) \ , \quad \text{para todo } i \neq j,$$

siendo $d(\mathbf{x}, \bar{\mathbf{x}}_i) = \log |\mathbf{S}_i| + (\mathbf{x} - \bar{\mathbf{x}}_i)' \mathbf{S}_i^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i)$ (véase el Problema 9.2 y, por ejemplo, Johnson y Wichern 2007 para una explicación más detallada). Implementamos este método mediante la siguiente función:

```
function [poblacion,D] = CuadratDiscrim3(x,X1,X2,X3)

% [poblacion,D] = CuadratDiscrim3(x,X1,X2,X3)
% Regla discriminante cuadratica para tres poblaciones
% Clasifica el individuo x en la poblacion 1, en la 2 o en la 3,
% utilizando la regla discriminante cuadratica.
% X1, X2 y X3 son muestras de las poblaciones 1, 2 y 3,
% respectivamente.
% D es el vector de distancias del individuo x a las
% poblaciones 1, 2 y 3.

x = x(:) ; px = length(x) ;
[n1,p1] = size(X1) ; [n2,p2] = size(X2) ; [n3,p3] = size(X3) ;
aux1 = [p1-p2, p2-p3, p1-p3] ;
if any(aux1 ~= 0)
    error('Las matrices de datos no tienen dimensiones coherentes')
else
    p = p1 ; clear p1 p2 p3
end
if px ~= p
    error('El vector x no tiene dimension adecuada')
else
    clear px
end

mgrande = [ mean(X1) ; mean(X2) ; mean(X3) ] ;
Sgrande = [ cov(X1) ; cov(X2) ; cov(X3) ] ;

D2 = zeros(3,1) ;
for i = 1:3
    Si = Sgrande([(i-1)*p+1:i*p],:) ;
    mi = mgrande(i,:) ;
    D(i,1) = log(det(Si)) + ((x'-mi) * inv(Si) * (x-mi')) ;
end
[Dmin,poblacion] = min(D) ;
```

Ahora los vectores `D` para las tres nuevas flores son:

ind.	$d(\mathbf{x}, \bar{\mathbf{x}}_1)$	$d(\mathbf{x}, \bar{\mathbf{x}}_2)$	$d(\mathbf{x}, \bar{\mathbf{x}}_3)$
$\mathbf{x}_1$	-2.0229	120.3326	187.5065
$\mathbf{x}_2$	441.8145	-7.1007	5.8556
$\mathbf{x}_3$	770.0216	4.0740	-5.7384

luego clasificamos estas flores en las mismas especies que habíamos determinado en el apartado (a).

PROBLEMA 9.6

Represéntense gráficamente los datos de las Tablas 8.1 y 9.1 en un diagrama de dispersión múltiple mediante la orden `gplotmatrix` *de Matlab, que permite diferenciar entre los distintos grupos. Esta orden sólo está disponible con la Statistics Toolbox.*

SOLUCIÓN

Consideremos primero los datos de la Tabla 9.1 que eran componentes de dos tipos de vino. Supongamos que tenemos los datos separados en dos matrices `X1` y `X2` correspondientes al vino 1 y al 2, respectivamente, como en el Problema 9.4. Para utilizar la orden `gplotmatrix` es necesario tener las observaciones en una única matriz, digamos `Datos`, y crear un elemento `Grupo` que contenga variables categóricas indicadoras del grupo al que pertenece la observación. El gráfico de la Figura 9.3 se puede crear con el siguiente código:

```
n1 = length(X1) ; n2 = length(X2) ;
Datos = Datos = [ X1 ; X2 ] ;
Grupo = cell(n1+n2,1) ;
for i=1:n1
   Grupo{i,1} = 'Vino 1' ;
end
for i=n1+1:n1+n2
   Grupo{i,1} = 'Vino 2' ;
end
gplotmatrix(Datos(:,1),Datos(:,2),Grupo,'kk','o*',[7 7],'on',...
            '','x_1','x_2')
```

Si hubiéramos definido `Grupo` como vector columna de la siguiente manera:

```
Grupo = [ ones(n1,1) ; 2*ones(n2,1) ] ;
```

en la leyenda sólo habrían aparecido los números 1 y 2. De ahí el haber creado el "cell array". Para representar gráficamente los datos de los iris procederemos de manera análoga. Consideramos las matrices de datos `X1`, `X2` y `X3` definidas en la solución del Problema 9.5 y utilizamos el código que sigue para obtener la Figura 9.4:

```
Datos = [ X1 ; X2 ; X3 ] ;
Grupo = cell(150,1) ;
for i=1:50
   Grupo{i,1} = 'I. setosa' ;
end
for i=51:100
   Grupo{i,1} = 'I. versicolor' ;
end
for i=101:150
   Grupo{i,1} = 'I. virginica' ;
end
Variables = cell(4,1) ;
Variables{1,1} = 'Long Sep' ; Variables{2,1} = 'Anch Sep' ;
Variables{3,1} = 'Long Pet' ; Variables{4,1} = 'Anch Pet' ;
gplotmatrix(Datos,Datos,Grupo,...
            'rbk','*o.',[],'on','',Variables(:,1),Variables(:,1))
```

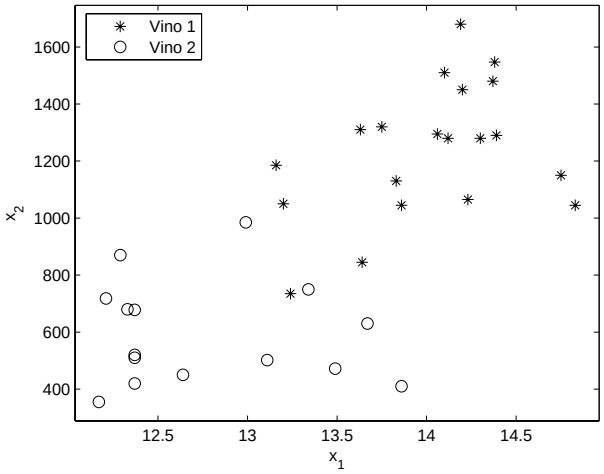


Figura 9.3.
Diagrama de dispersión de los datos de vinos (Problema 9.6)

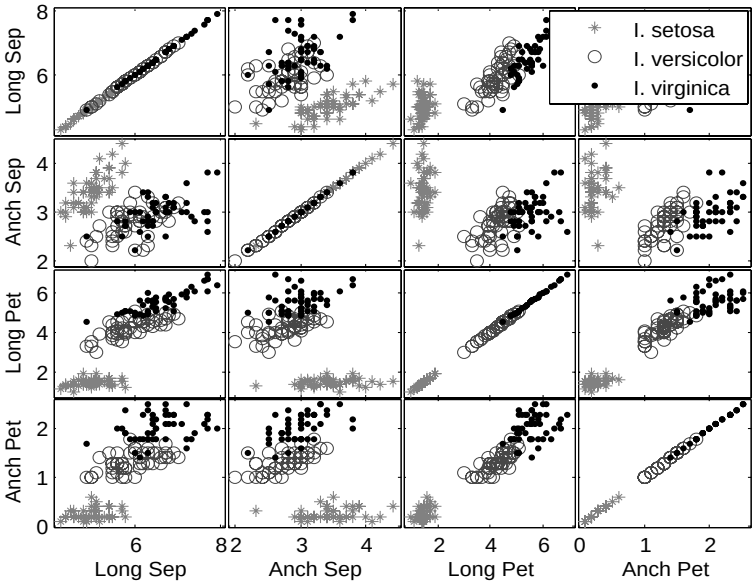


Figura 9.4.
Diagrama de dispersión de los datos de iris (Problema 9.6)

PROBLEMA 9.7

*Podemos estimar la tasa de error de una regla de clasificación mediante un procedimiento de validación cruzada (*cross-validation*) propuesto por Lachenbruch y Mickey (1968), que describimos a continuación para el caso de dos poblaciones.*

Paso 1. *Comenzar con las observaciones de la población 1, x_{1i}, $i = 1, \ldots, n_1$. Apartar una observación x_{1i} de la muestra y construir una regla de clasificación con las restantes $n_1 - 1$ observaciones de la población 1 y los n_2 datos de la población 2.*

Paso 2. *Clasificar el dato x_{1i} utilizando la regla construida en el Paso 1.*

Paso 3. *Repetir los Pasos 1 y 2 hasta que se hayan clasificado todas las observaciones de la población 1. Calcular m_1, el número de observaciones de la población 1 mal clasificadas.*

Paso 4. *Repetir los Pasos 1 a 3 para las observaciones de la población 2. Denotar por m_2 el número de observaciones de esta población mal clasificadas.*

Prográmese la anterior secuencia de pasos en Matlab para los datos del Problema 9.4 y la regla discriminante lineal. Estímese $P(i|j)$, la probabilidad de clasificar erróneamente en la población i una observación que en realidad proviene de la población j, mediante $\hat{P}(i|j) = m_j/n_j$. Estímese también la tasa global de error mediante $(m_1 + m_2)/(n_1 + n_2)$. Dibújense los datos en un gráfico de dispersión y señálese cuáles son los que están mal clasificados.

SOLUCIÓN

Suponemos ya introducidas en Matlab las matrices `X1` y `X2` con los datos de las poblaciones 1 y 2, respectivamente. A continuación escribimos la función que estima la probabilidad de clasificación errónea y la tasa global de error. Por ejemplo, el valor de $\hat{P}(2|1)$ lo da `EC1`. Se utiliza la función `LinealDiscrim` del Problema 9.4.

```
function [EC1,EC2,TGE] = TasaErrorDiscLin(X1,X2)

% TasaErrorDiscLin
% Estimacion de la tasa de error en la regla discriminante lineal
% con dos poblaciones con muestras X1 y X2.
% Devuelve:
% EC1 = probabilidad de clasificar en poblacion 2 un dato de
%       poblacion 1;
% EC2 = probabilidad de clasificar en poblacion 1 un dato de
%       poblacion 2;
% TGE = tasa global de error de clasificacion.

[n1,p] = size(X1) ; [n2,p] = size(X2) ;
```

```
Poblacion1 = zeros(n1,1) ; Poblacion2 = zeros(n2,1) ;

for i =1:n1
    if i == 1
        X1menos1 = X1([2:end],:) ;
    else
        X1menos1 = X1([1:i-1,i+1:n1],:) ;
    end
    Poblacion1(i) = (LinealDiscrim(X1(i,:),X1menos1,X2)~=1) ;
end

for i =1:n2
    if i == 1
        X2menos1 = X2([2:end],:) ;
    else
        X2menos1 = X2([1:i-1,i+1:n2],:) ;
    end
    Poblacion2(i) = (LinealDiscrim(X2(i,:),X1,X2menos1)~=2) ;
end
m1 = sum(Poblacion1); m2 = sum(Poblacion2) ;
EC1 = m1/n1 ; EC2 = m2/n2 ; TGE = (m1+m2)/(n1+n2) ;
```

Obtenemos las estimaciones $\hat{P}(1|2) = 0$ y $\hat{P}(2|1) = 0.1$.
El diagrama de dispersión de los datos aparece en la Figura 9.5. Hemos marcado con flechas los datos mal clasificados de la población 1 (del vino 2 no hay ninguno). Para ello hemos utilizado esencialmente las órdenes:

```
aux1 = find(Poblacion1) ;
MalClasif1 = X1(aux1,:) ;
```

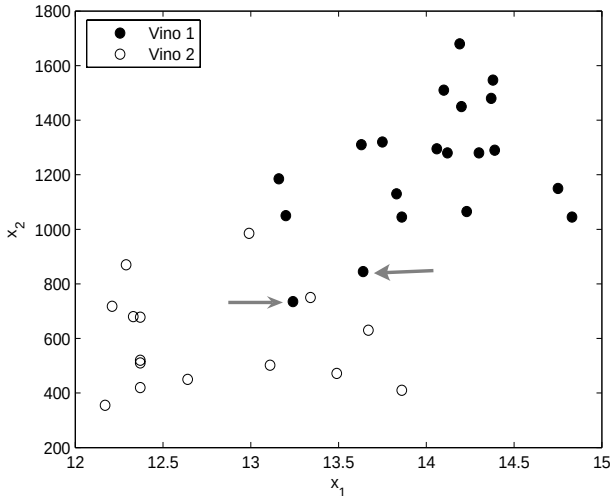


Figura 9.5.
Diagrama de dispersión y datos mal clasificados (Problema 9.7)

PROBLEMA 9.8

Para los datos de la tabla 8.1, estímese la tasa de error cometida con las reglas discriminantes propuestas en el Problema 9.5.

SOLUCIÓN

Escribimos sólo la función de Matlab que hace el cálculo para el caso de la regla discriminante lineal. Para el caso de la regla cuadrática es totalmente análogo.

```
function [EC1,EC2,EC3,TGE] = TasaErrorDiscLin3(X1,X2,X3)

% TasaErrorDiscLin3
% Estimacion de la tasa de error en la regla discriminante lineal
% para tres poblaciones con muestras X1, X2 y X3.
% Devuelve
% EC1 = probabilidad de clasificar mal un dato de la poblacion 1;
% EC2 = probabilidad de clasificar mal un dato de la poblacion 2;
% EC3 = probabilidad de clasificar mal un dato de la poblacion 3;
% TGE = tasa global de error.

[n1,p] = size(X1) ; [n2,p] = size(X2) ; [n3,p] = size(X3) ;
Poblacion1 = zeros(n1,1) ; Poblacion2 = zeros(n2,1) ;
Poblacion3 = zeros(n3,1) ;

for i =1:n1
    if i == 1
        X1menos1 = X1([2:end],:) ;
    else
        X1menos1 = X1([1:i-1,i+1:n1],:) ;
    end
    Poblacion1(i) = (LinealDiscrim3(X1(i,:),X1menos1,X2,X3)~=1) ;
end

for i =1:n2
    if i == 1
        X2menos1 = X2([2:end],:) ;
    else
        X2menos1 = X2([1:i-1,i+1:n2],:) ;
    end
    Poblacion2(i) = (LinealDiscrim3(X2(i,:),X1,X2menos1,X3)~=2) ;
end

for i =1:n3
    if i == 1
        X3menos1 = X3([2:end],:) ;
    else
        X3menos1 = X3([1:i-1,i+1:n3],:) ;
    end
    Poblacion3(i) = (LinealDiscrim3(X3(i,:),X1,X2,X3menos1)~=3) ;
end
```

```
m1 = sum(Poblacion1); m2 = sum(Poblacion2) ; m3 = sum(Poblacion3) ;
EC1 = m1/n1 ; EC2 = m2/n2 ; EC3 = m3/n3 ;
TGE = (m1+m2+m3)/(n1+n2+n3) ;
```

PROBLEMA 9.9

Los datos de la Tabla 9.2, extraídos de Newman et al. (1998), son observaciones tomadas sobre pacientes que han sufrido un ataque al corazón. Las variables consideradas son X_1 la edad a la que el paciente sufrió el ataque, X_2 y X_3 sendas medidas de contractilidad del corazón, X_4 la dimensión ventricular izquierda al final de la diástole y X_5 una medida de cómo se mueven los segmentos del ventrículo izquierdo. La clase 0 está constituida por aquellos pacientes que sobrevivieron menos de un año desde el ataque. La clase 1 son los que sí sobrevivieron.
Se tienen observaciones referentes a dos pacientes nuevos:

Paciente	X_1	X_2	X_3	X_4	X_5
1	*70*	*0.173*	*16.02*	*5.20*	*18.56*
2	*62*	*0.224*	*12.45*	*4.71*	*14.38*

*y se desea clasificarlos en alguna de las dos poblaciones. Para ello se utiliza la regla k-NN (k-*nearest neighbours*) o de los k vecinos más próximos. Dada una observación $\mathbf{x}$ a clasificar, se toman las k observaciones $\mathbf{x}_i$ de la muestra más cercanas a $\mathbf{x}$. Se clasifica $\mathbf{x}$ según el "voto de la mayoría", es decir, se asigna $\mathbf{x}$ a la clase 0 si el número de k-vecinos que pertenecen a esta clase es mayor que el de los que pertenecen a la clase 1.*
Impleméntese en Matlab la regla k-NN para dos poblaciones y utilícese con $k = 5$ para clasificar a los nuevos pacientes.
Observación: *Para ilustrar el método, utilizamos la distancia euclídea como medida de proximidad entre observaciones. Dependiendo de la naturaleza de los datos sería conveniente reemplazarla por alguna de las distancias propuestas en el Capítulo 5.*

SOLUCIÓN

Podemos utilizar el siguiente código, fácilmente generalizable a mayor número de poblaciones. La matriz `Datos` está formada por las cinco primeras columnas de la Tabla 9.2 y el vector `Clase` por la última columna de esta tabla. `NuevaObs` es el vector de observaciones correspondiente a un nuevo paciente.

```
function ClaseNuevaObs = kNNClasif(NuevaObs,Datos,Clase,k)

% kNNClasif(NuevaObs,Datos,Clase,k)
% Clasifica NuevaObs utilizando la regla k-NN (k vecinos mas
```

Tabla 9.2.
Enfermos del corazón (Problema 9.9). Fuente: Newman *et al.* (1998)

X_1	X_2	X_3	X_4	X_5	Clase
71	0.260	9.000	4.600	14.00	0
72	0.380	6.000	4.100	14.00	0
55	0.260	4.000	3.420	14.00	0
60	0.253	12.062	4.603	16.00	0
57	0.160	22.000	5.750	18.00	0
68	0.260	5.000	4.310	12.00	0
62	0.230	31.000	5.430	22.50	0
60	0.330	8.000	5.250	14.00	0
46	0.340	0.000	5.090	16.00	0
54	0.140	13.000	4.490	15.50	0
77	0.130	16.000	4.230	18.00	1
62	0.450	9.000	3.600	16.00	0
73	0.330	6.000	4.000	14.00	0
60	0.150	10.000	3.730	14.00	0
62	0.120	23.000	5.800	11.67	1
55	0.250	12.063	4.290	14.00	0
69	0.260	11.000	4.650	18.00	1
62	0.070	20.000	5.200	24.00	1
66	0.090	17.000	5.819	8.00	0
66	0.220	15.000	5.400	27.00	1
69	0.150	12.000	5.390	19.50	1
85	0.180	19.000	5.460	13.83	1
73	0.230	12.733	6.060	7.50	1
71	0.170	0.000	4.650	8.00	1
55	0.210	4.200	4.160	14.00	0
61	0.610	13.100	4.070	13.00	0
54	0.350	9.300	3.630	11.00	0
70	0.270	4.700	4.490	22.00	0
79	0.150	17.500	4.270	13.00	0
59	0.030	21.300	6.290	17.00	0
58	0.300	9.400	3.490	14.00	0
60	0.010	24.600	5.650	39.00	1
66	0.290	15.600	6.150	14.00	0
63	0.150	13.000	4.570	13.00	0
57	0.130	18.600	4.370	12.33	0
70	0.100	9.800	5.300	23.00	0
79	0.170	11.900	5.150	10.50	0
72	0.187	12.000	5.020	13.00	0
51	0.160	13.200	5.260	11.00	0
70	0.250	9.700	5.570	5.50	0
65	0.360	8.800	5.780	12.00	0
78	0.060	16.100	5.620	13.67	0
86	0.225	12.200	5.200	24.00	1
56	0.250	11.000	4.720	11.00	0
60	0.120	10.200	4.310	15.00	0
59	0.290	7.500	4.750	13.00	0
54	0.217	17.900	4.540	16.50	0
64	0.200	7.100	4.580	14.00	0
54	0.070	16.800	4.160	18.00	0
78	0.050	10.000	4.440	15.00	1
55	0.280	5.500	4.480	22.00	0
59	0.344	9.100	4.040	9.00	0
74	0.200	4.800	4.560	12.50	0
65	0.160	8.500	5.470	16.00	1
58	0.170	28.900	6.730	26.08	1
70	0.380	0.000	4.550	10.00	0
63	0.300	6.900	3.520	18.16	1
59	0.170	14.300	5.490	13.50	0
57	0.228	9.700	4.290	11.00	0
78	0.230	40.000	6.230	14.00	1
62	0.260	7.600	4.420	14.00	1

```
% proximos).
% Variables de entrada:
%   NuevaObs = vector a clasificar con numero de componentes p
%   Datos = Matriz de datos nxp con individuos de
%   clase (0 o 1) conocida.
%   Clase = Vector nx1 con etiquetas 0 o 1 de los individuos de
```

```
%   la muestra.
%   k = Numero de vecinos mas proximos a NuevaObs para su
%   clasificacion.
% Variable de salida:
%   Clase=0 (resp. 1) si la mayoria de los k-NN son de
%   clase 0 (resp. 1).
%   En caso de empate se sortea Clase aleatoriamente.

% Control del numero de variables de entrada
if nargin < 4
    error('Faltan variables de entrada')
end

NuevaObs = NuevaObs(:) ; % "Obligamos" a NuevaObs a que sea
                         %   vector columna
Clase = Clase(:) ;

% Control de la dimension de variables de entrada
[n,p] = size(Datos) ; p2 = length(NuevaObs) ;
[nC,pC] = size(Clase) ;
if n ~= nC
    error('El numero de filas de la muestra no coincide con...
           el de la clase')
end
if p ~= p2
    error('El numero de datos de la nueva observacion no es...
           coherente con la dimension de la muestra')
end
if pC ~= 1
    error('La clase tiene que ser un vector, no una matriz')
end
clear nC pC p2

% Calculamos la distancia euclidea de NuevaObs a la muestra
DistEuclid = sum((Datos - ones(n,1) * NuevaObs').^2,2) ;
[DistEOrd,IndEOrd] = sort(DistEuclid) ; % Ordenamos las distancias.
ClasekNN = Clase(IndEOrd([1:k])) ; % Clases de los k-NN
NumkNN1 = sum(ClasekNN == 1) ; % Numero de kNN en Clase 1.
NumkNN0 = sum(ClasekNN == 0) ; % Numero de kNN en Clase 0.
if NumkNN1 > NumkNN0
    ClaseNuevaObs = 1 ;
elseif NumkNN1 < NumkNN0
    ClaseNuevaObs = 0 ;
else % Se "tira una moneda al aire" y se decide la clase
     %  aleatoriamente
    u = rand(1,1) ;
    ClaseNuevaObs = (u >= 0.5) ;
end
```

Tomando $k = 5$ asignaremos el paciente 1 a la clase de los que sobrevivirán más de un año y el paciente 2 a la otra clase.

PROBLEMA 9.10

Generalícense las funciones Matlab escritas en los Problemas 9.4 y 9.5 para implementar la regla discriminante lineal con un número genérico g de clases. Aplíquese la nueva función para clasificar los dos pacientes del Problema 9.9.

Indicación: *La nueva función detectará el número g de poblaciones entre las que hay que discriminar como la longitud del vector $\mathbf{n} = (n_1, n_2, \ldots, n_g)'$, que contiene los tamaños muestrales n_i observados en cada población i. El vector $\mathbf{n}$ será una variable de entrada de la función.*

SOLUCIÓN

Proponemos el siguiente código

```
function poblacion = LinealDiscrimg(x,X,vector_n)

% LinealDiscrimg(x,X,vector_n)
% Regla discriminante lineal para cualquier numero g de
% poblaciones
% Variables de entrada:
%   x Observacion a clasificar: vector de p componentes
%   vector_n Vector de dimension gx1, que contiene
%            n1, n2, ..., ng, siendo
%            ni el numero de observaciones en la poblacion i, para
%            i=1,2,...,g.
%   X Matriz de datos de dimension nxp, con n=n1+n2+...+ng,
%   que contiene las matrices de datos X1, X2, ..., Xg de las
%   poblaciones puestas en orden una encima de la otra.
x = x(:) ; px = length(x) ;
g = length(vector_n) ;
[n,p] = size(X) ;

if px ~= p
    error('La dimension de x no es coherente con la de la...
           matriz de datos X')
else
    clear px
end

mMatriz = zeros(g,p) ; S = zeros(p) ;

for i=1:g
    if i ~= 1
        Fila1 = sum(vector_n([1:i-1]))+1 ;
        Filani = Fila1 -1 + vector_n(i) ;
    else
        Fila1 = 1 ; Filani = vector_n(1) ;
    end
    Xi = X([Fila1:Filani],:) ;
    mMatriz(i,:) = mean(Xi) ;
```

```
    Si = cov(Xi) ;
    S = S + (vector_n(i)-1) * Si ;
end

S = S/(sum(vector_n)-g) ;
x_rep = (ones(g,1) * x') - mMatriz ;
D2 = diag( x_rep * inv(S) * x_rep' ) ;
[D2min,poblacion] = min(D2) ;
```

Para aplicar esta función a los datos del Problema 9.9 utilizaremos las mismas matrices `Datos` y `Clase` que usábamos en este problema. Con el código

```
IndVivos = find(Clase==1) ;   IndMuertos = find(Clase==0) ;
vector_n = [length(IndMuertos) ; length(IndVivos)];
X = [Datos(IndVivos,:) ; Datos(IndMuertos,:)];
x = [ 70 ; 0.173 ; 16.02 ; 5.20 ; 18.56 ] ;
poblacion = LinealDiscrimg(x,X,vector_n)
```

comprobamos que la clasificación de los nuevos pacientes coincide en este caso con la de la regla k-NN.

PROBLEMA 9.11

Estímese la tasa de error cometida con la regla k-NN en el Problema 9.9 mediante el procedimiento de validación cruzada (véase el Problema 9.7). Calcúlese dicha estimación para $k = 1, 2, \ldots, n-1$, siendo n el tamaño muestral total. Decídase si el valor de k que minimiza la tasa global de error es el más adecuado. En caso contrario, propóngase otro método de elección del número de vecinos.

SOLUCIÓN

Utilizaremos las matrices `Datos` y `Clase` del Problema 9.9. El código que estima la tasa de error mediante validación cruzada para k vecinos es el siguiente:

```
function [EC1,EC0,TGE] = TasaErrorDisckNN(Datos,Clase,k)

% TasaErrorDisckNN
% Estimacion de tasa de error en regla kNN con
% dos poblaciones (0 y 1).
% Variables de salida:
%   EC1 = probabilidad de clasificar en Poblacion 0
%         un dato de la 1.
%   EC0 = probabilidad de clasificar en Poblacion 1
%         un dato de la 0.
%   TGE = tasa global de error.
```

```
[n,p] = size(Datos) ;
ErrorSi = zeros(n,1) ;

for i =1:n
  if i == 1
    DatMenos1 = Datos([2:n],:) ; ClMenos1 = Clase([2:n],:) ;
    ErrorSi(1) = ( kNNClasif(Datos(1,:),DatMenos1,...
                  ClMenos1,k) ~= Clase(1)) ;
  elseif i == n
    DatMenos1 = Datos([1:n-1],:) ; ClMenos1 = Clase([1:n-1],:) ;
    ErrorSi(n) = ( kNNClasif(Datos(n,:),DatMenos1,...
                  ClMenos1,k) ~= Clase(n)) ;
  else
    DatMenos1 = Datos([1:i-1,i+1:n],:) ;
    ClMenos1 = Clase([1:i-1,i+1:n],:) ;
    ErrorSi(i) = ( kNNClasif(Datos(i,:),DatMenos1,...
                  ClMenos1,k) ~= Clase(i)) ;
  end
end

n1 = sum(Clase == 1) ; n0 = sum(Clase == 0) ;
m1 = sum((Clase == 1).*ErrorSi) ;
m0 = sum((Clase == 0).*ErrorSi) ;
EC1 = m1/n1 ; EC0 = m0/n0 ;
TGE = (m0+m1)/(n0+n1) ;
```

A continuación calculamos la tasa de error para $k = 1, \ldots, n-1$:

```
TasaError = zeros(n-1,3) ;
for k =1:n-1
    [EC1,EC0,TGE] = TasaErrorDisckNN(Datos,Clase,k) ;
    TasaError(k,:) = [EC1,EC0,TGE] ;
end
```

y obtenemos las tres primeras columnas de la Tabla 9.3. Aparecen sólo los resultados para $k \leq 25$ porque para $k \geq 26$ el resultado coincide con el de $k = 25$. La menor tasa global de error corresponde a $k = 20$ o $k = 21$, pero esta elección no sería adecuada, ya que probablemente erraríamos en la clasificación de observaciones de la población 1. Un mejor procedimiento en la elección del número de vecinos es tomar aquel valor de k que minimice la suma de cuadrados $EC1^2$+$EC0^2$ (véase la Tabla 9.3) y que en este caso sería $k = 2$ o $k = 4$.

Tabla 9.3.
Tasa estimada de error en regla k-NN (Problema 9.11)

k	EC1	EC0	TGE	$EC1^2+EC0^2$
1	0.6471	0.1591	0.2951	0.4440
2	0.6471	0.1364	0.2787	0.4373
3	0.7059	0.1364	0.2951	0.5169
4	0.6471	0.1364	0.2787	0.4373
5	0.8235	0.0909	0.2951	0.6865
6	0.8824	0.0909	0.3115	0.7868
7	0.8824	0.0909	0.3115	0.7868
8	0.7647	0.0909	0.2787	0.5930
9	0.9412	0.0682	0.3115	0.8905
10	1	0.0909	0.3443	1.0083
11	1	0.0682	0.3279	1.0046
12	1	0.0227	0.2951	1.0005
13	1	0	0.2787	1
14	1	0	0.2787	1
15	1	0	0.2787	1
16	1	0	0.2787	1
17	1	0	0.2787	1
18	1	0	0.2787	1
19	1	0	0.2787	1
20	0.9412	0	0.2623	0.8858
21	0.9412	0	0.2623	0.8858
22	1	0	0.2787	1
23	1	0	0.2787	1
24	1	0	0.2787	1
25	1	0	0.2787	1

Referencias

Cuadras, C. M. (2004). Análisis multivariante. Manuscrito accesible en http://www.ub.es/stat/personal/cuadras/cuad.html.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179–188.

Frets, G. (1921). Heredity of head form in man. *Genetica* 3, 193–400.

Gower, J. (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika* 55, 582–585.

Hutcheson, M. y Meyer, M. (1996). DASL The Data and Story Library. http://lib.stat.cmu.edu/DASL/DataArchive.html.

Iordansky, N. (1973). The skull of the Crocodilia. In C. Gans y T. S. Parsons (Eds.), *Biology of the Reptilia, Vol. 4*, New York, pp. 201–262. Academic Press.

Johnson, R. A. y Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Prentice Hall.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23, 187–200.

Lachenbruch, P. A. y Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* 10, 1–11.

Mardia, K. V., Kent, J. T. y Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.

Nash, W. K., Sellers, T. L., Talbot, S. R., Cawthorn, A. J. y Ford, W. B. (1994). UCI Repository of machine learning databases. University of California, Irvine. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Newman, D. J., Hettich, S., Blake, C. L. y Merz, C. J. (1998). UCI Repository of machine learning databases. University of California, Irvine. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Peña, D. (2002). *Análisis de Datos Multivariantes*. McGraw-Hill.

Índice de funciones y código Matlab

Índice de conceptos